

Summers' winter of discontent

The president of Harvard has learnt a painful lesson in public communication. The media and Harvard academics may have over-reacted to his comments about women in science, but there is an opportunity to benefit from the affair.

To judge from media reports last week, you would think that Harvard University's president Larry Summers had told everyone in a skirt to quit science, pack up their handbag and head home to the kids — clearly biologically incapable of rivalling men in this intellectually thorny field.

What Summers actually said has been harder to discern. There is no transcript of the speech he made at an academic meeting. But by most accounts, he talked about potential reasons why so few women reach top positions in science and engineering. And he at least raised the possibility that innate differences could play a part.

Whatever his actual words, the interpretation — that women suffer an inborn and insurmountable intellectual handicap in science — whipped up a firestorm of protest both on and off Harvard's (male-dominated) campus. Swamped by fuming letters and stinging media reports, Summers has released serial cringing apologies in which he emphasized his efforts to bump up the number of female scholars at Harvard — actions for which some staff, at least, give him credit.

His latest remarks are consistent with his reputation for making blunt and provocative comments. That hasn't gone down well at Harvard, where his radical plans for change have raised hackles since he took over in 2001 (see *Nature* **433**, 190–193; 2005).

The latest debate about gender inequality has provided an opportunity to look afresh at the causes and potential solutions. Is there any evidence for innate differences in the sexes' cognitive abilities? Researchers say there is. One well-explored area is gender disparity in

processing spatial problems: men and women have been shown to use different areas of the brain to weave their way out of a virtual-reality maze (G. Grön *et al.* *Nature Neurosci.* **3**, 404–408; 2000).

Unfortunately, this doesn't resolve the question in hand. Men's and women's brains may work in slightly different ways, but researchers cannot say for sure whether these disparities underlie differences in their exam performance or ability to scoop top academic jobs. By the time adolescents take tests or women are applying for tenure, it is difficult to tease apart possible contributory biological differences from the complex cocktail of other factors that could be holding them back. These include society's sexual stereotyping, dismal family support and old-fashioned discrimination. Summers also mentioned some of these in his speech.

The Harvard president needs to think more carefully before he speaks, even when, as here, the occasion is off the record. But the hullabaloo should not deter academics from discussing important if disquieting questions, both on principle and also in the hope of pinpointing causes and finding solutions.

The faculty at Harvard and other institutions should view this furore as an opportunity to highlight shortfalls in the system that handicap women, and demand that they be improved. Female scientists, engineers and mathematicians at Harvard might find this a good moment to ask for a promotion — and those outside its hallowed walls should mail in an application. With some concerted action, the fallout could be a boost for female scientists everywhere. ■

Data sharing for disasters

Last month's tsunami and its aftermath have highlighted a need for more science — and more effective sharing of data.

At last week's United Nations World Conference on Disaster Reduction in Kobe, Japan, space agencies discussing the applications of remote sensing received a lashing. They went there to highlight successes in using satellite images for disaster monitoring and rescue activities (such as December's tsunami), and to discuss the further development of their technologies. Researchers enthused about the use of satellite images to help dispatch rescue teams. For those whose houses were washed away, high-resolution 'before' and 'after' shots could even be used to claim property boundaries.

Impressive stuff, but is more high technology the answer? Remote sensing has so far not lived up to the promise that was attributed to it in the heady days when the projects were planned. The science has improved, but the dissemination of data has not.

Then there are the practical obstacles to using satellite data in the places where they are most needed — underdeveloped areas that have been hit by natural or other disasters. For example, wind speed and other cyclone-related data come in differently formatted tables and with different units, making them difficult to use in a crisis situation. These problems are not insurmountable but could make all the difference in a crisis.

To their credit, space-agency representatives recognized the problems and almost unanimously saw a need to push ahead with getting

more people using the data. But it is not clear how to move forward, with each side expecting the other to do more. Non-profit middleman companies, such as UK-based MapAction, can provide maps, and did so in Sri Lanka. Another solution would be to include in space-agency grants a requirement for local public outreach activities.

Claims that more high-tech solutions are required are questionable. Some proposals, such as expanding the use of satellites to give world-wide, real-time coverage of ocean surfaces, smack of scientific opportunism in the face of disaster — this must be avoided. Many of the problems that beset the Indian Ocean countries were not due to a lack of technological hardware. Many of the buoys, tide gauges and seismic stations that were in place were not put to proper use, partly because countries did not share effectively (see page 343).

Before the international scientific community sets out to right the wrongs of disaster preparedness and recovery, there needs to be an evaluation of what is already there and how it can be better used. New equipment to come on line in the Indian Ocean needs to be integrated to get the best results. Space agencies need to make sure that their satellite data get to the users. Countries on the Indian Ocean and elsewhere need to be encouraged to share data from their tide gauges and seismic stations as widely as possible.

In short, researchers must get data to where they are needed. ■

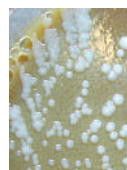




Ground zero
Geologists fear
survey funds
may dry up
p342



Warning signs
Lack of unity
hinders tsunami
early-alert projects
p343



Dishing dirt
Bacteria scandal
harms US plans
for biodefence lab
p344



Morale boost
Bill Gates and
Norway give funds
for vaccinations
p345

Medicare compels heart patients to enlist in follow-up research

Meredith Wadman, Washington

Many elderly Americans may find that the government will only pay for expensive drugs and devices if they agree to enrol in long-term follow-up studies to gather data on safety and efficacy.

Medicare, the publicly funded \$320-billion US health-insurance scheme for the elderly and disabled, announced this week that it will supply a costly pacemaker-like device to hundreds of thousands of people at risk from cardiac arrest who previously did not qualify for the implant.

But the offer, which could cost Medicare \$3 billion annually, has strings attached. Applicants must be included in a national registry that will require data, for example, on the brand of device implanted, how often it fires, and disease progression.

Some also see the move as blurring various national and international ethical rules. These typically forbid anyone to require that patients participate in research as a condition of receiving medical care and stipulate that subjects must give free and informed consent when taking part in such research.

The follow-up study is research, argues Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, Philadelphia. He points out that the data collected are for epidemiological purposes that do not directly benefit the individual patient. "You are supposed to be free to participate in research without coercion, without any form of 'You do this or else,'" he says, "and this gets closer to the 'or else'."

But Medicare officials argue that monitoring the long-term effectiveness of treatments is an economic necessity, given the soaring costs of new drugs and devices, combined with an ageing population, particularly where data on the clinical efficacy of treatments are incomplete.

More systematic trials of this sort could also yield important safety data, points out Sean Tunis, chief medical officer at the Center for Medicare and Medicaid Services (CMS) in Baltimore, Maryland, which decides what treatments Medicare covers. He cites as evidence the painkiller rofecoxib, marketed as Vioxx, which was on the market for five years



Data to hand: Medicare will fund pacemaker-like implants only for those who join a national registry.

before it was found to increase the risk of heart attacks and strokes in long-term users. "There is a missing piece of the clinical research enterprise, a body of clinical and scientific questions that doesn't get answered in the current universe of studies," says Tunis.

"Developing better evidence through data registries and practical clinical trials is an increasingly important feature of Medicare's coverage decisions," Tunis wrote in last week's *New England Journal of Medicine* (352, 222–224; 2005), with co-author Mark McClellan, administrator of the CMS and former chief of the US Food and Drug Administration. Indeed, Medicare indicates that it will extend this approach to other costly drugs and devices. And private insurance companies are likely to follow its lead.

Impulsive action

Under its new plan, Medicare will increase the number of patients considered for implantable cardioverter-defibrillators (ICDs) to around 500,000, but individual patients must agree to take part in the study before they become eligible.

ICDs are battery-powered devices that, like pacemakers, are implanted in the chest to deliver electrical impulses to the heart. But unlike pacemakers, which fire repeatedly to

correct a slow heartbeat, an ICD fires only when it detects rapid, irregular beats.

Until now, Medicare has paid for ICDs only in patients with particular types of heart failure, or those who have previously experienced dangerous heart arrhythmias. The change of plan was prompted by a major study published last week (*N. Engl. J. Med.* 352, 225–237; 2005), which showed that the devices are more effective in preventing death than either a placebo or the antiarrhythmic drug amiodarone. So, increased use of ICDs could cut deaths from congestive heart failure.

Information technology now allows data to be collected more easily and cheaply than ever before, point out Tunis and McClellan. Even so, doctors will have to agree to collect and input the required data, and some are sceptical of the practicality of a register.

But many in the academic and clinical communities welcome the data-gathering requirement. "Most of us think that it's a fantastic move," says Ezekiel Emanuel, who chairs the department of clinical bioethics at the National Institutes of Health. "I do think it raises a question of when research and clinical care merge. But my own sense is that every patient ought to be a data point, because we would learn so much more about what does and doesn't work in a clinical setting."

NASA urged to lay plans for mission to Europa

Tony Reichhardt, Washington

US planetary scientists are urging NASA to make a conventional mission to Jupiter's moon Europa a top priority.

The call has been spurred by increasing doubts surrounding the agency's plan to send a nuclear-powered craft to Europa and two of Jupiter's other satellites, Ganymede and Callisto.

The troubled Jupiter Icy Moons Orbiter (JIMO) is to be NASA's first nuclear-powered mission, but its launch date has already slipped from 2011 to 2015, largely because of the novel technology involved.

Sensing trouble, a scientific advisory group for JIMO last month wrote to Andrew Dantzler, NASA's chief of Solar System exploration, to advise the agency to consider a conventionally powered Europa mission "if major slips beyond an expected JIMO launch in the 2015 time frame occur".

Along with Mars, Europa is seen as the most likely place in the Solar System to find extraterrestrial life. Robert Pappalardo, a planetary scientist at the University of Colorado in Boulder, is among some 80 scientists who have signed a white paper, attached to the JIMO advisory group's report, emphasizing the importance of Europa exploration. "We want to remind NASA that with or without JIMO, Europa is a top priority," Pappalardo says.

NASA surprised many scientists in 2003 by proposing JIMO as the first mission under its Project Prometheus nuclear-propulsion programme.

But JIMO has run into such problems (see *Nature* 431, 113; 2004) that NASA has ordered a comprehensive review of Prometheus, which should be complete in April. One option under consideration is to defer the Jupiter probe and begin the project with a smaller, more manageable nuclear mission, perhaps to our own Moon or to an asteroid.

No one is ready to declare JIMO dead just yet, but NASA is realizing that reaching Jupiter is "quite a big first step" for its nuclear programme, says Dantzler. He adds that he has already commissioned an informal study into a conventional mission to Europa. "I'm not putting a lot of money into it yet," he says. "But if it does come down to a conventional Europa mission, I want to hit the ground with the wheels already turning."



Ground force: researchers take readings from an EarthScope seismic station in Socorro, New Mexico.

Cash shortfall threatens to rock US geophysics project

Rex Dalton, San Diego

US Earth scientists fear that lack of funds could derail a megaproject to examine the geophysical structure of North America.

EarthScope, funded by the National Science Foundation (NSF), is designed to generate a three-dimensional view of the geophysical processes affecting Earth's mantle, tens of kilometres below the continent's surface. It should revolutionize scientists' understanding of volcanoes, fault systems and earthquakes.

Construction of the project's infrastructure is now into its second year, and involves some \$200-million worth of equipment. Although Congress authorized the NSF to buy and install the equipment over five years, it has delayed a decision on funding for the operation and maintenance — estimated at \$13 million per year — and the \$100 million required for research over the first decade of operations.

Approved in 2003, when the NSF's \$4.2-billion annual research budget was expected to grow rapidly, EarthScope now faces a starkly different financial climate — the NSF's budget for the 2006 fiscal year is expected to be flat or even slightly reduced.

On 24 January, EarthScope submitted a 500-page research plan for the coming years, including budget requests for operation and maintenance. EarthScope officials declined to discuss financial specifics, except to say that they were in line with earlier projections.

But there is no clear picture of the near-term finances for individual research projects. Just \$4 million was available for EarthScope research this year, officials say. "People now are very nervous about the whole thing," says Adam Dziewonski, a

seismologist at Harvard University who chairs a seven-strong advisory panel to the USArray, one of the project's components.

"The Earth science community's concern is merited. We need to work together to identify all available funding for research and operations," says the NSF's programme director for EarthScope, Kaye Shedlock.

EarthScope has three main components. The USArray is a mobile system involving more than 800 seismometers that will move east from California across the nation for a decade, visualizing the geophysical structure. The Plate Boundary Observatory (PBO) is a system of devices for examining tectonic plate interactions from Alaska to the lower states. And the San Andreas Fault Observatory at Depth (SAFOD) experiment will drill a core into the coastal mountains south of San Francisco, installing pressure gauges and seismometers in the active fault zone to record movements and quakes.

USArray seismometers are now being deployed in California. SAFOD scientists have drilled more than 2.5 km deep, and plan to core into the fault this summer. PBO researchers are also preparing to set out equipment in Alaska this summer.

Gregory van der Vink, EarthScope's project director, acknowledges that there is concern in the geoscience community, but he is confident that the programme can secure the necessary research funds in the coming budget discussions. The goal is not just to put instruments in the ground and data in an archive, but to use them in research, van der Vink says. "We will do everything we can to win funds for research. This is the data set for the next generation of geoscientists," he adds.

Solo efforts hamper tsunami warning system

David Cyranoski, Kobe

A plan to set up a tsunami warning system for the Indian Ocean — and eventually the whole world — received enthusiastic support in Kobe, Japan, last week. But observers cautioned that the job is being made harder by a lack of coordination and data sharing between countries.

At the United Nations World Conference on Disaster Reduction, scientists and policy-makers from around the globe were quick to pledge national contributions to the concept.

So far, the most concrete plans for hardware to support a warning system have been laid out by India. It has pledged to spend US\$30 million from February to expand its network of tide gauges, seismic stations and 'tsunameters' — ocean-bottom monitors that can detect changes in pressure from a mere one-centimetre-high wave.

"We have a long coast line and a great number of natural disasters besides tsunamis. We want this system very badly," says S. K. Subramanian, head of the India Meteorological Department in New Delhi. "When we issue a warning, it will be based on the Indian system."

In the meantime, Germany plans to spend some US\$40 million adding ten 'new generation' tsunameters and 40 seismic stations in and around Indonesia. And the United States has announced a plan to spend \$37.5 million expanding its Pacific warning system (see 'A system that works ... if people listen', below).

Coordination is necessary both to ensure



Chasing the waves: deep-ocean buoys could be marshalled for a global tsunami warning system.

that resources are not wasted and to avoid potential confusion from conflicting warnings, says Laura Kong, director of the International Tsunami Information Center in Honolulu, Hawaii. "Instruments in the Indian Ocean are currently owned and used by a number of countries. We must ensure coordination and sharing of data," she says.

But so far coordination has been seen as lacking. One US representative, for example, noted at the conference that he had learned more about Germany's technical plans by

talking to the press than from meetings with German representatives themselves.

UNESCO's Intergovernmental Oceanographic Commission will host two technical meetings in Paris in March to discuss how to move forward in light of the many initiatives that are emerging — although this will be after some of the efforts, including India's, have begun.

India says that its system, which should be up and running in 2007, will offer warnings to the entire region. But Kong argues that without one internationally designated centre to coordinate the information, confusion could arise. "What will countries in the region do when they get different estimates and different warnings?" she asks.

India has also been criticized for not sharing tide-gauge data that are essential for understanding the ocean's dynamics, and for refusing access to some researchers keen to study the country and its islands in the wake of the earthquake.

James Whitcomb of the US National Science Foundation's Earth-science division noted at the meeting that easier access to India's seismic data could have cut the time it took for scientists to realize the full scale of the 26 December earthquake — which was originally estimated at 8.0, but was later upgraded to 9.0. Seismic data from north of the source area — from India or China, for example — would have helped, said Whitcomb.

Indian representatives pointed out that their data are available on a website (www.imd.ernet.in) within half an hour of an event, but Whitcomb said that real-time data, as offered by many other countries, would be preferable for global warnings. ■

A system that works ... if people listen

As representatives from around the world gathered in Kobe, Japan, to discuss earthquakes and tsunamis, a magnitude 6.8 earthquake hit the sea some 200 km southeast of Tokyo. Within minutes, Japan's warning system had kicked in, broadcasting a message on television screens across the country saying that the quake would not cause dangerous waves for those on land, but that there was limited danger for those in the water.

Most tsunami experts agree that nowhere is more prepared for tsunamis than Japan, which gets hit by damaging waves roughly once every seven years. "It's the best-prepared nation both in terms of hardware and general awareness," says Laura Kong, director of the International Tsunami Information Center in Honolulu, Hawaii.

Earthquakes are detected within two minutes by seismic monitors all around the country. Seismic information is quickly fed into a computer simulation that contains 100,000 pre-calculated outcomes for earthquakes of various magnitudes and depths at 4,000 different locations. One minute later, one of six regional

centres, staffed around the clock, can disseminate warnings by TV or radio.

For tsunamis triggered by earthquakes farther away, Japan is helped by the Pacific Tsunami Warning Center in Hawaii — a US system consisting of six seafloor pressure sensors — which triggers alarms for Pacific nations. This is the only oceanic tsunami warning system currently in place. It has won plaudits for detecting tsunamis and evaluating the danger — it was successfully used to cancel a tsunami warning and prevent an evacuation in Hawaii in November 2003, for example. But it isn't perfect: three of its six buoys have been out of commission since at least September 2004; one has been down for 18 months. The United States plans to upgrade the system to 38 buoys across the Pacific, Atlantic and Caribbean by mid-2007.

But despite all this investment, Pacific nations still face the problem of human error. Tohoku University tsunami specialist Fumihiko Imamura notes that even when tsunami warnings are issued in Japan, people go to the beach to watch.

David Cyranoski

Europe pares down double patents on breast-cancer gene

Alison Abbott, Munich

In battles worthy of a TV court drama, the European Patent Office (EPO) has slashed the scope of two patents on the breast-cancer gene *BRCA1*. Mutations in this gene predispose women to some hereditary forms of breast cancer.

The outcome can be seen as a victory for those who use the gene to diagnose cancers, but the rulings also raise an ethical conundrum. The gene's most common mutation — frequent in a single group, the Ashkenazi Jews — remains protected by patents; 33 others do not.

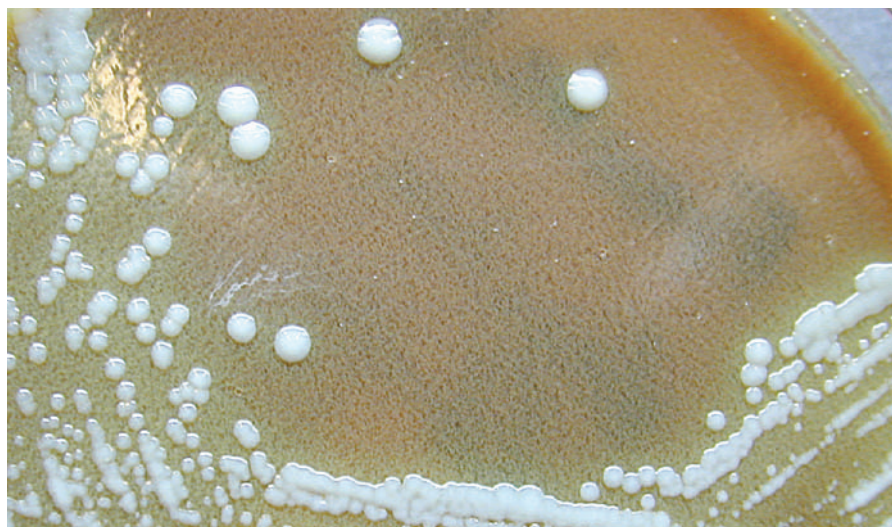
In the 1990s, several research groups raced against each other to find the sequence of *BRCA1*. Myriad Genetics of Salt Lake City in Utah rushed to file a flurry of patents — three of them in Europe — based on the sequence. Myriad required all tests for the gene to be carried out in its labs in Salt Lake City, prompting opposition from enraged clinical researchers in Europe.

Last May, the EPO revoked in its entirety one of Myriad's three patents. In its rush to secure rights, Myriad had filed its sequence of the normal gene with the US patent office in a rough form. It updated the filed sequence regularly, but the EPO ruled that the correct sequence had been filed only after other scientists had deposited the correct sequence in a public database, making it invalid for patenting (see *Nature* 429, 329; 2004).

Myriad has filed an appeal against this decision. It has also transferred ownership of the European patents to the University of Utah, which is now handling opposition to the other two patents, one covering the gene product itself and the other covering specific mutations.

On 20 January the EPO ruled that the patent on the gene product — whose scope originally included any probe, or nucleotide sequence, that can recognize the gene — should be restricted to just one probe. This probe corresponds to a part of the gene whose sequence was completely correct in an early filing.

A ruling on the third patent came on 25 January. This patent originally covered 34 known mutations of *BRCA1*, but had been restricted, during negotiations in advance of the hearing, to a single mutation — the most common one, which occurs with highest frequency in Ashkenazi Jews. The hearing supported the patent, overruling objections that finding a common mutation in a known gene is not truly inventive. ■



No one is sure how an active strain of tularaemia bacteria managed to infect university lab workers.

Infection scare inflames fight against biodefence network

Rex Dalton, San Diego

The accidental infection of three Boston University researchers with dangerous bacteria is raising questions about the US expansion of biodefence labs.

Massachusetts-based Boston University has been chosen by the National Institutes of Health to host one of two major research and containment centres for deadly pathogens. The university is planning to build a National Biocontainment Laboratory in a well populated area of south Boston, much to the dismay of residents. The US\$128-million facility will be part of a system of labs being developed across the United States.

But last year, experiments by researchers at Boston University's medical campus caused three of them to get tularaemia infections. "This is symbolic of what can happen," says Boston University epidemiologist David Ozonoff, who switched last year from supporting the biodefence lab to opposing it.

The researchers thought they were handling the tularaemia bacterium, which could be used for bioterrorism, in its deactivated form. In May, two of the researchers became ill with respiratory infections. No one suspected the cause until October, after a third researcher had sickened. The US Centers for Disease Control and Prevention, based in Atlanta, Georgia, was notified about the cases in mid-November.

Since then, the patients have recovered. But the handling of the case has caused outrage and stirred opposition to the biodefence lab. The university delayed reporting the incidents to state health authorities by 12 days, although state law requires that such cases be reported within 24 hours. And the university — along with city, state and federal officials

— did not publicly disclose the infections until last week. This meant the issue was not discussed last autumn during the planning process for the biodefence lab.

"If there was any risk outside these three infected people, we would have made a general announcement in a heartbeat," says Thomas Moore, acting provost of Boston University's medical campus, pointing out that tularaemia cannot be passed from person to person. He says that announcements were delayed until after internal investigations, and adds that the infections were not considered relevant to public debate over the biodefence lab, because that will have better containment facilities.

In the aftermath of the incident, Boston University's chief of infectious disease at the time, Peter Rice, was removed from his position. And Boston health officials say they are preparing an education campaign for lab personnel, to ensure that safety regulations are followed. It is thought that the researchers who became ill may have been working without the required fume hood.

But details of the case continue to raise questions. For instance, Paul Mead, an epidemiologist at the Centers for Disease Control and Prevention, says his agency has been unable to determine how and where the non-infectious sample became contaminated with a wild, infectious strain. "This is a bit of a freak event," says Mead. Perhaps tularaemia-contaminated rabbit blood was used as a culture medium, he says. Genetic testing of material may pin down the source.

Officials at the National Institute of Allergy and Infectious Diseases, which will fund the Boston biodefence lab and the national system, declined to comment on the incident as an investigation is still under way. ■

Global vaccine project gets a shot in the arm

Erika Check, Washington

Two donors this week gave US\$1 billion to a project that immunizes children in the world's poorest countries against preventable diseases such as hepatitis B and yellow fever.

On 24 January, the Seattle-based Bill and Melinda Gates Foundation and the Norwegian government announced ten-year grants of \$750 million and \$290 million, respectively, to the Global Alliance for Vaccines and Immunization (GAVI).

That is a big sum, but, to put it in context, a further \$8–10 billion will be needed over the next ten years just to get existing vaccines out to the children who need them, according to the World Health Organization (WHO). And that is before counting the extra costs of delivering new vaccines, such as shots against rotavirus and pneumococcus, as these come on to the market.

The sum of \$10 billion is not much more than the international community has pledged to tsunami relief in three weeks. However, “the tsunami was a very visible demonstration of need in developing countries”, says David Fleming, director of global health strategies for the Gates foundation. “The issue with immunization is that millions of children are dying needlessly, but because deaths do not occur in a single disaster, it is harder to attract the attention of donors.”

Geneva-based GAVI, whose partners include the United Nations' children's fund, the WHO, governments and commercial vaccine-makers, was created in 1999 with \$750 million of Gates funding, and other donors have since brought its total funding to \$1.3 billion. “We think this is the best investment we've ever made,” said Bill Gates in a teleconference with reporters, pointing



Booster shot: billions of dollars are still needed to get jabs to children in poorer countries.

out that GAVI programmes have prevented the deaths of 670,000 children born between 2001 and 2003 in 71 countries.

“We boosted routine immunization coverage for diphtheria, tetanus and whooping cough from 53% in 2000 to 81% in 2003,” said Yoweri Museveni, president of Uganda, in a statement. “And because GAVI has a unique incentive system, our funding actually increased when we reached our immunization targets.”

As well as delivering vaccines, the large sums involved are indirectly encouraging more vaccine research, says Julian Lob-Levyt,

executive secretary of GAVI. “Assurance of markets over the long-term encourages industry to invest greater resources in research and development for urgently needed vaccines, such as those for malaria.”

But much remains to be done. In 2002, 1.4 million children under five years of age died from vaccine-preventable diseases. In 2003, more than 27 million children went without vaccinations in their first year of life. Attacking this problem is key to reducing child mortality — one of the UN Millennium Development Goals, a series of commitments signed by all UN members in 2000. The target is to reduce the under-five mortality rate by two-thirds between 1990 and 2015, but so far the world looks as if it will need an extra two decades to achieve this.

Global health is on the agenda at the annual meeting of the World Economic Forum being held this week in Davos, Switzerland. Gates will take his place on a panel with Tony Blair, the UK prime minister, and the outspoken rock star Bono. Blair's government is pushing for an International Finance Facility that would raise \$50 billion a year in development aid between now and 2015: the facility would sell bonds on capital markets that would be recovered by donor funds pledged over several decades. Gates says he is very excited about this approach.

“Modest amounts of resources can save large numbers of lives,” says Gates, regretting that making global health an international spending priority is “very, very difficult”.

“When people see tsunami victims on the TV, they do want to reach out; they do want to provide resources and help,” he adds. “We all need to think creatively about how we can tap into the human compassion that's there.” ■

MIT wraps up Dublin lab following funding failure

Roxanne Khamisi, London

Media Lab Europe, the European arm of the world-famous Media Lab at the Massachusetts Institute of Technology (MIT) in Cambridge, is to close next week, less than five years after it opened.

The research lab, based in Dublin, was launched in 2000 during the dotcom bubble. It was a joint venture between MIT and the Irish government, which invested €35 million (US\$46 million) in the centre. The original plan was that corporate sponsors would supplement the cost of running the lab, which has some 60 staff, including around 40 scientists. But in 2003 the lab spent more than €8 million and raised only €2.6 million. Overall, it managed to hook just eight partners. “It was difficult to attract the



Media Lab Europe is closing its doors.

corporate funding we had hoped for,” says Sorcha Duggan, a spokeswoman for the lab. In 2004, the lab asked the government to fund its next three years with a €9-million grant, but the request was refused.

The stock-market bubble burst in 2000, slashing the share value of potential

sponsors. Observers also argue that the lab's research, which included projects on “the future of human relationships as mediated by technology”, was too abstract to attract investment, especially in Europe, which has less of a culture of corporate sponsorship than the United States.

“It's a bit of a blow,” says Seán Duke, editor of *Technology Ireland*. He points out that Ireland has taken pride in its ability to attract foreign high-tech investment.

MIT pulled out of its Media Lab Asia in New Delhi in 2003. This followed a row with the Indian government, which wanted the futuristic centre to focus on more practical goals (see *Nature* 423, 213; 2003). The latest closure leaves Media Lab with no links to spin-offs outside the United States. ■

Duo bids to create climate of change at nuclear-weapons lab

Washington Two anti-nuclear organizations have announced their intention to bid together for management of the Los Alamos National Laboratory in New Mexico.

If selected, the consortium members say, they would redirect the nuclear-weapons research facility away from weapons activity and towards the study of climate change, alternative energy sources, and the environmental clean-up of the surrounding land. The proposal is a long shot, they admit, but they hope the move will influence the final selection.

Los Alamos — one of the nation's three nuclear-weapons labs — has been managed by the University of California for more than 60 years. But a string of recent security lapses has led the US Department of Energy, which oversees the lab, to open the contract to competition (see *Nature* **423**, 104; 2003).

The new bidders are made up of Tri-Valley Communities Against a Radioactive Environment, a watchdog group based in Livermore, California, and Nuclear Watch of New Mexico, a Santa Fe-based body that advocates nuclear disarmament.

The terms of the competition are

expected to be finalized next month, and bidders will probably have to announce their intentions soon after. Several partnerships between universities and private companies are thought likely to compete.

Botanic gardens provide safe haven for rare plants

London More than 9,000 of the world's endangered plant species are being cultivated successfully in botanic gardens,



Blooming rare: botanic gardens are home to endangered plants such as the titan arum.

according to a survey by UK charity Botanic Gardens Conservation International.

The results of the two-year study, released on 18 January, show that about a quarter of the world's known endangered plants have found refuge in the 500 gardens surveyed. Some 34,000 plant species are known to be endangered, but up to 100,000 plants are thought to be under serious threat of extinction from habitat destruction and global climate change. The gardens provide a seed bank of endangered plants, but conservationists say this is no replacement for healthy wild stocks.

Japan warms to joint climate study with UK

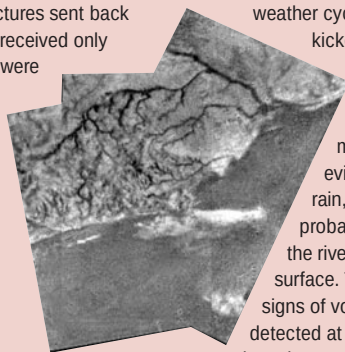
Tokyo Britain and Japan last week initiated a five-year collaboration to study climate change using Japan's Earth Simulator, one of the world's most powerful supercomputers. Japan has been criticized in the past for not releasing computer time for international use.

Six scientists from the NCAS Centre for Global Atmospheric Modelling in Reading and the Met Office's Hadley Centre will be stationed in Yokohama to use the computer. They plan to run their climate models with a much better resolution than is currently possible. This will allow them to take into account smaller-scale phenomena, such as

Huygens keeps half an eye on Titan's weather

Munich Scientists analysing pictures sent back from Saturn's moon Titan have received only half the number of photos they were hoping for from Huygens, the European Space Agency (ESA) probe that landed on Titan on 14 January. About 350 snaps taken by the probe were not received because ESA controllers neglected to send a computer command telling the right hardware to switch on.

Despite this, photos from Huygens (right), along with other data, have provided clues to the moon's



weather cycle. The probe's landing kicked up some methane gas from the ground, which is made largely of water ice and organic molecules. This provides evidence of recent methane rain, scientists say, which probably also helped to carve the river channels seen on the surface. The mission also found signs of volcanic activity: elements detected at the surface suggest that the volcanoes spewed out water and ammonia, rather than lava.

wind or water eddies, which can affect the climate. They also hope to develop a model that incorporates chemical and biological processes and their effects on climate.

Gender bias unveiled as big problem in Indian science

New Delhi Sexual harassment and gender discrimination are top-priority problems for female scientists in India, according to a study funded by the Indian National Science Academy (INSA). The report, conducted by a nine-member panel of independent

scientists (seven of them women), concludes that male-centred policies are key obstacles for female scientists.

Women seldom get awards or fellowships, it points out. No woman in the past five years has received a Bhatnagar award, India's top science prize, of which ten were awarded to young scientists this year. In 50 years, the Council of Scientific and Industrial Research has not appointed a woman as director of any of its 42 labs. INSA has never had a female president in its 70-year history, and all its top officials except one have been male.

INSA is setting up a committee to decide

how to implement the recommendations, which include the creation of a central grievance agency, more support through facilities such as crèches, and the inclusion of more women in decision-making bodies.

Brain donation scheme probed in consent row

Washington The collection of about 100 human brains in Maine for a prominent research institution in Maryland is under investigation, after questions were raised about whether proper consent was given for the donations.

A federal attorney is now interviewing families to decide whether there were any legal violations when these brains were collected, between 1999 and 2003, for the Stanley Medical Research Institute in Bethesda. Several families have come forward to say they did not give permission for the entire brain to be used in research, and the institute has settled one lawsuit with regard to such a claim.

The institute, which focuses on mental diseases such as schizophrenia and bipolar disorder, provides tissue samples and funds to laboratories for research. It says that it followed all appropriate laws in securing the organs. The brains were collected for Stanley by an embalmer in Maine.



JIMIN LAI/AF/GETTY IMAGES

It was 8:45 on a tropical December morning when Mohamed Nazfer saw the tsunami coming. He remembers the time exactly because he was eating breakfast, listening to the news on the radio. When he heard screaming in the street outside his house on the beachfront of Kalmunai, a fishing village on the east coast of Sri Lanka, he walked a few steps towards the beach to see what all the fuss was about. Then he began to run, fast, from the white wall of water hurtling towards him.

Although the 35-year-old fisherman survived the Indian Ocean tsunami, his home, livelihood, wife, ten relatives and some 6,000 other people here and in neighbouring communities did not. Sometimes, Nazfer tells me as we stand among the wreckage, he thinks that he would be better off dead. He lives in a refugee camp now, a short distance inland, but he comes down every day to roam the remnants of his home. He remembers hearing somewhere that the waves were triggered by an earthquake 1,500 kilometres away, but he doesn't much care about the science behind the disaster. "It was God," he says.

For the scientists who have spent their lives researching the physics of tsunamis, however, there is great reason to care. Philip Liu, a fluid dynamicist and tsunami expert at

On the trail of destruction

Quirin Schiermeier travels to Sri Lanka with a team of scientists in the wake of last month's tsunami. Together with locals they search through the damage for clues of where the wave hit hardest.

Cornell University in Ithaca, New York, is one such person. In the aftermath of the disaster, he could be found hastily organizing an expedition to Sri Lanka. He and his team wanted to measure the traces that the flood had left on the soil, vegetation and structures in the 'inundation zone'.

I joined the ten-strong group of coastal engineers, geologists and fluid mechanics from the United States, New Zealand and Sri Lanka as they set out to collect data to determine the height, velocity and extent of the flood at different stretches of the coast. With

additional, more detailed information about the local topography and the near-shore ocean depths, they hope to improve numerical tsunami models, and to produce physical maps that local planners can use when rebuilding settlements, hotels and infrastructure. Those same maps could also be used to plan better evacuation routes, and to save lives.

The 26 December tsunami was the first to strike Sri Lanka in living memory. Across the Indian Ocean, tsunamis are rare but not unknown. In June 1994, a tsunami killed

TIMELINE 26.12.04

00:59 GMT

An earthquake of magnitude 9.0 strikes off the west coast of Sumatra. The sea floor moves several metres upwards, sparking a tsunami.

3 minutes

Seismic signals reach Cocos Island Station in the Indian Ocean, a facility that is part of a US-led university consortium.

8 minutes

Seismic signals from the quake are received at the Pacific Tsunami Warning Center (PTWC) in Hawaii from sensors in Australia. An alarm is triggered.

15 minutes

The PTWC issues a bulletin stating a quake of magnitude 8.0 has occurred, with no risk of tsunamis to Pacific nations. The US Geological Survey (USGS) circulates information about the quake to internal officials.

some 200 coastal residents in east Java, after an earthquake occurred off the coast. Waves were reported to have been as high as 15 metres. No one knows for sure, but it has been estimated that large tsunamis may hit this part of the world once every 1,000 years.

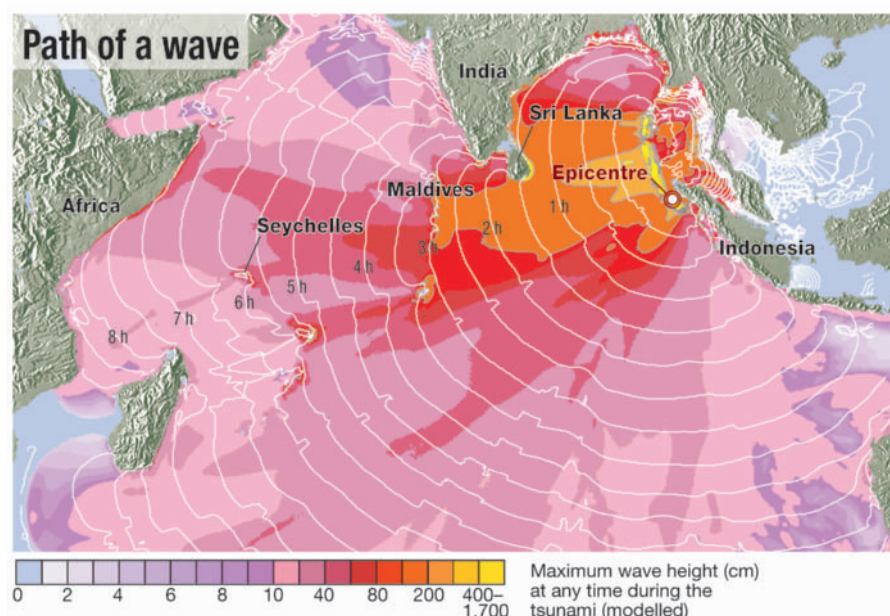
The recent tsunami was one of the world's worst natural disasters in decades. About 33,000 people, almost half of whom were children, died on Sri Lanka's east and south coasts alone. The death toll around the entire Indian Ocean region exceeds 220,000, and numbers are still rising. Although other earthquakes, such as that in Tangshan, China, in 1976, have taken hundreds of thousands of lives, no single incident has had such a devastating effect on so many countries for a long time — and certainly not with the overriding feeling that something could have been done to save many of the victims. Early warning systems against tsunamis do exist, but at present only for the Pacific region, where the waves occur much more frequently (see page 343). The information about this wave took hours to reach the people who could have acted (see Timeline, below).

Tough challenge

Researchers such as Liu know that their efforts should ultimately help improve the warning systems. But collecting the necessary information is no easy task. Liu's expedition on 10–15 January proved to be a first-rate logistical challenge. Plans and schedules had to be changed on a daily basis, as roads proved impassable, bridges were destroyed and hotels closed. But thanks to the team's improvisational skills and much support from the Sri Lankan military, who acted as hosts, it did provide a wealth of data.

Once in Sri Lanka, Liu's team split in two — with journalists in tow. One group, including Liu and myself, set off on 10 January from Colombo to Trincomalee, a small town on the east coast. From there, we headed south, visiting affected coastlines down to Amparai, a former stronghold of the Tamil rebels. A second team, led by Costas Synolakis, a civil engineer and tsunami expert at the University of Southern California in Los Angeles, surveyed the better-developed south coast of the island.

The destruction became increasingly severe the farther south we travelled, often



with seemingly random differences in effects from one beach to another. The 50,000-strong town of Trincomalee suffered comparatively modest damage, but the small communities on the neighbouring Koddiiyar Bay were hit hard. South of Batticaloa things deteriorated rapidly, until finally we came to the shocking sight of Kalmunai. Here, the clean-up had just begun, and the smell of decay still lingered in the air — when we arrived some local volunteers had just buried a human head on the beach.

"It is a conflicting experience for me to go to the field amid all this devastation," admits Bruce Jaffe, a geological oceanographer with the Pacific Science Center in Santa Cruz, California. "In a way it makes you feel guilty. But then these data are really invaluable for better assessing the hazard and frequency of tsunamis."

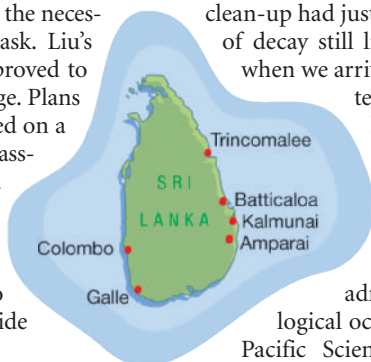
On each beach, Jaffe scanned the ground, the rooftops and even empty swimming pools for sediment deposits left behind by the wave. He also recorded signs of the erosion that the rushing water seemed to have caused hundreds of metres inland. He took sand samples from different sediment layers and packed them in hundreds of carefully labelled plastic bags for analysis back in the lab. From the thickness of the deposits and

the distribution of grain size in the tsunami-generated layers, he will later use computer models to back-calculate the number of incoming waves, the flow velocities and water heights at every probed location.

Sifting the clues

"Post-tsunami surveys help us define what kind of resolution we need in our simulations," says Synolakis, who has taken part in 12 such field trips since 1992. As well as Jaffe's approach of looking at grain size, others hope to use different measures to test and calibrate their models. Physicists such as Liu tend to use evidence of high-water marks to back-calculate wave heights and speeds. He and his colleagues used visual clues to determine the direction of the incoming flow on Sri Lankan beaches. Local children, who crowded round us and tried to steal the tools from our hands, helped to seek out buildings with clear water lines and posed for pictures. Despite the tragedy of last month, they were intensely curious and keen to help.

Between the daily forays onto the beaches, the researchers spent hours in the car, with laptops bouncing on their knees, debating the specifics of this wave. Tsunami research is done by a small community: no more than 200 scientists worldwide specialize in the field, says Liu. And many of them are now scouring the beaches of Sri Lanka, Indonesia and the other countries hit by the wave. Late into the night, members of our



25 minutes

The USGS revises magnitude to 8.2.

30 minutes

Sumatra hit by the tsunami.

1 hour

The PTWC issues second bulletin upgrading the quake to magnitude 8.5 and identifies the possibility of a tsunami near the epicentre. The centre's team attempts to notify colleagues in Indonesia without success.

2 hours

Eastern Sri Lanka hit by the tsunami. Several satellites pass over the Bay of Bengal and record information about tsunami wave heights. Scientists do not get the data until hours later.

team animatedly discussed what the wave looked like at birth, and how it propagated across the Indian Ocean.

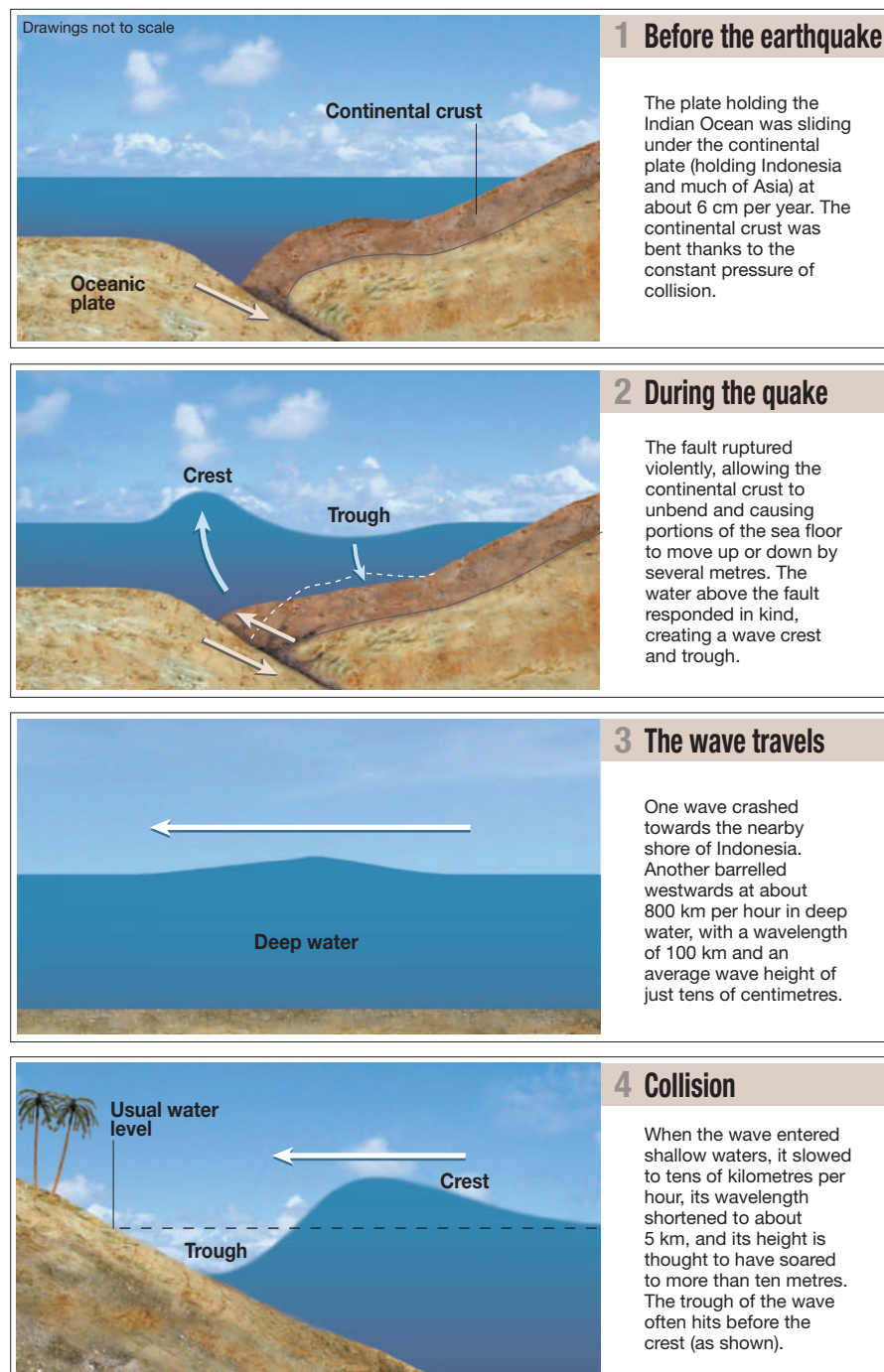
The mechanism by which tsunamis form is relatively simple (see Graphic, right). And the equations that describe their rough characteristics, such as speed and wave height, in deep water are similarly straightforward. But there are complexities that are far more difficult to deal with. December's tsunami was caused when an underwater fault line ruptured across hundreds of kilometres — and no one is sure what a wave formed in this way actually looks like.

Rapid rise

Most of the experts agree that a tsunami is initially shaped like an 'N', with a large crest and trough (see 'During the quake', right), before evolving into more complicated shapes. If the trough of the wave travels ahead, as it seems to have done towards Indonesia, the ocean recedes from the shore line before the wave crest arrives. This phenomenon led many eyewitnesses of the tsunami to wander down to the beach out of curiosity — leaving them little or no time to escape the full force of the incoming wave.

After birth, a tsunami propagates in deep water with stunning speed but with a low wave height. In this case, the wave moved at about 800 kilometres per hour. Models put the maximum wave height to the east of Sri Lanka at more than a metre (see Map, previous page), whereas satellites measured up to 80 centimetres. Aboard a ship, such waves would be hardly perceptible. The small size also means it is extremely difficult to detect tsunamis from the air or with buoys that measure variations in water pressure. Filtering out the loud 'noise' from normal wind-generated waves on top of a long tsunami hump requires sophisticated equations. It took hours to process and analyse the data gathered by satellite for the December wave. Such systems are not designed for, nor suitable for, warnings — by the time the results were ready the tsunami had already struck.

Once tsunami waves reach shallow water, they transform dramatically. Because the speed of a tsunami is a function of the water depth, it slows as it nears the shore. But because its energy remains almost constant, the height of the wave grows tremendously in shallow water. On our trip, we saw one 5-metre-high building that had been completely submerged. Reefs, bays and undersea



features such as canyons all play a role in modifying the wave, buffering or focusing its energy, or changing its direction. Reflected waves interact with each other. And the steeper the sub-sea slope, the more quickly the wave grows and the more energy it retains when it hits land.

This aspect of a tsunami's journey is incredibly difficult to model. "Things start to get really complex once the waves hit the

shore and continue to flow on land," says Liu. Topography, land material, friction and numerous small-scale features from trees to houses come into play — creating a modelling nightmare.

Attempts to describe the breaking, turbulence and debris pick-up of a tsunami on land are still very crude and need a lot of development, says Patrick Lynett, a coastal modeller at Texas A&M University in College Station,

2 hours 30 minutes

Internet newswire reports of casualties in Sri Lanka emerge. India hit by the tsunami.

3 hours

The US ambassador in Sri Lanka sets up system to notify the prime minister in case of any large aftershocks, based on information from the PTWC.

3 hours 22 minutes

Aftershock of magnitude 7.1 strikes off the Nicobar Islands. (More than a dozen aftershocks of magnitude 6.0 or greater occurred — this was the largest.)

3 hours 30 minutes

Maldives hit.

who has modelled the 1998 Papua New Guinea tsunami¹. “Comparison with the real thing is of paramount importance,” he says.

A tsunami’s final surge onto land rarely resembles the towering waves often drawn by artists. The waves rarely ‘break’ the way that surfing waves do. Rather the water simply rushes inland, often preceded by a ‘bore’ — a mass of extremely turbulent white water. What such a bore looks like exactly was also a matter for midnight debates on this trip — not one of the tsunami researchers has actually seen an incoming tsunami.

Instead, fluid dynamicists study the propagation and transformation of waves under simplified laboratory conditions, using large water tanks equipped with wave-making mechanisms and advanced optical-measurement technology.

A rush and a push

The resulting mathematical equations are then marshalled to create computer-generated tsunamis, using software such as the Tunami-N2. This standard model was designed in the early 1990s by Nobuo Shuto, a modelling expert at Tohoku University’s Disaster Control Research Center in Sendai, Japan, to calculate wave physics up to a shoreline. With a similar model — the Cornell Multigrid Coupled Tsunami model — Liu has simulated the propagation and wave evolution of the Indian Ocean tsunami.

But these models have limitations. They are based on an idealized ocean floor, and on untested assumptions made about the seismic generation of tsunamis in the source region. Such lab waves are not made by anything approaching a rupturing fault. And there are issues about how water behaves on different scales that cannot be resolved in a lab model. Real-life tsunamis consistently catch these models out. A wave that hit the Indonesian island of Flores in 1992, for example, defied standard models. To explain its behaviour, Liu had to create a more detailed simulation², which showed that the tsunami could bend around corners and penetrate into sheltered coastal areas without losing its energy.

The accuracy of models has improved considerably during the past decade, says Synolakis. “At the time of the 1992 Nicaragua tsunami, models underestimated the observed flooding by a factor of ten,” he says. “Since then, the difference between models and measurements has diminished to around a factor of two.”



Tracking the wave: Sri Lankan children help to measure the high water mark (above) while geologists extract tsunami deposits (below).



Preliminary data suggest that this time the models may have done even better. Vasily Titov, a former student of Synolakis, now at the Pacific Marine Environmental Laboratory in Seattle, Washington, used a model³ that simulates how a tsunami wave penetrates inland. The results seem to agree with measurements made so far to within 30%.

To close that gap further, the researchers will continue to collect data. They are also pressing for more accurate maps of the sea floor around Sri Lanka. At a concluding meeting on 15 January in Colombo, Sri Lankan navy officials said that they would provide high-resolution maps of the sea floor immediately, and help to collect any missing information. Eventually, says Synolakis, it might be possible to produce inundation maps of threatened coastlines in Sri Lanka, showing

how far water is likely to intrude inland under various threats. Such maps currently exist only for some coastal areas in Japan and the United States. The results are unlikely to persuade fishermen to move inland, he concedes, but they will help town planners to rebuild destroyed schools, hospitals and tourist infrastructure at secure locations.

Structurally unsound

The information can’t come soon enough: at many places reconstruction has already begun, but is often substandard. Many builders are using bad, crumbling concrete and little steel to reinforce the structures — much like most of the buildings that were destroyed by the wave. “Some hastily rebuilt constructions look like they may collapse without a tsunami,” says Synolakis.

The geological work being done by Jaffe’s team will also help planners to assess the risk of future events. By digging deeper into coastlines or marshes, where sediments have been preserved for thousands of years, Jaffe hopes to show how often large tsunamis have occurred in this region — a method that has previously worked for Papua New Guinea and the west coast of the United States⁴. A mythological account in the *Mahavamsa*, Sri Lanka’s national Buddhist chronicle, suggests that a tsunami may have hit the island around 150 BC. This book reports that “the sea flooded the land, as a punishment by God”, says Starin Fernando, a geologist at Sri Lanka’s Geological Survey and Mines Bureau in Colombo. Using Jaffe’s methods, Fernando is hoping to find geological evidence for this religious tale.

Such work may reaffirm that tsunamis in this area are extremely rare. But the lessons learned about modelling by Liu and his team will have far wider implications. They now have reams of valuable data that will improve models of how tsunamis behave near shore and on land. And next time, they hope, all this information will help to curtail the extent of any disaster.

Quirin Schiermeier is Nature’s German correspondent.

1. Lynett, P., Borrero, J. C., Liu, P. L.-F. & Synolakis, C. E. *Pure Appl. Geophys.* **160**, 2119–2146 (2003).
2. Yeh, H., Liu, P., Briggs, M. & Synolakis, C. *Nature* **372**, 353–355 (1994).
3. Titov, V. V. & Synolakis, C. E. *J. Waterway Port Coast. Ocean Eng.* **124**, 157–171 (1998).
4. Jaffe, B. E. & Gelfenbaum, G. in *Solutions to Coastal Disasters ’02* 836–847 (American Society of Civil Engineers, Reston, VA, 2002).

For more on the tsunami visit:

♦ www.nature.com/news/specials/tsunami

4 hours 25 minutes

Team at Harvard University’s seismology department upgrades the original quake to 8.9 using a technique that analyses both the size and shape of seismic waves. The USGS is due to get this same technology this year.

7 hours 15 minutes

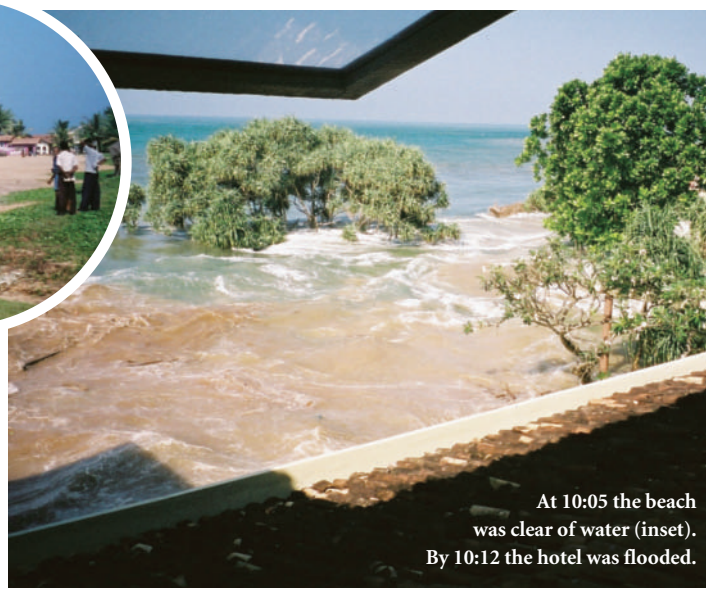
The PTWC advises the US state department about potential threat to Madagascar and Africa. Africa is hit around the same time.

14 hours 30 minutes

The PTWC warns that waves have reached the Pacific, and may continue for a day or two. Water from the tsunami reaches as high as 2.6 metres above the normal water mark in Mexico — probably due to focusing of energy by a seafloor ridge in the Pacific — and half a metre on both the east and west coasts of North America.

Get off the beach — now!

As a theoretical seismologist, Chris Chapman says he “sits in front of the computer trying to work out how seismic waves can tell us things about the interior of the Earth”. Sitting in Ahungalla, 30 kilometres north of hard-hit Galle on the southwest tip of Sri Lanka when last month’s tsunami rolled in, he got a fresh perspective on earthquakes and their impact. Fortunately, Chapman, who works at Schlumberger Cambridge Research in Britain, was able to surmise the danger and initiate efforts to clear the beaches. He tells David Cyranoski about the experience.



At 10:05 the beach was clear of water (inset).
By 10:12 the hotel was flooded.

PICTURES: C. CHAPMAN

a nearby plate boundary, and that this first wave was all that was going to happen.

Nevertheless she went off and spoke to the hotel manager and warned him that there had been an earthquake and there might be more to come. She’s Canadian, and she’s always found English understatement difficult to understand.

When did you know that things would get worse?

At 9:50 the sea was far below its normal level. That was about when I said to my wife that this was going to be something big. She is more outgoing than me, so she rushed back to the hotel manager and told him to get people off the beach. The sea level was continuing to go down. By 10:05 it was several metres below normal.

What kind of response did you get?

The hotel manager was extremely good. As soon as my wife said what might happen, he got his staff out to the beach with megaphones and told people to come in. By the time the sea level started to rise again, we were all near the hotel; not necessarily on higher floors, but within running distance of the stairs.

There was no panic until about 10:10, at which point the water was coming in very quickly. The striking thing was not the wave but the speed and sheer volume of water. It was amazing. The water took 35 minutes to retreat, but only 7 minutes to come back in. Fortunately no one from our hotel was lost.

How high was the wave in the end?

I didn’t see it because I was running for my life. It’s difficult to estimate heights, but the total rise was 5 to 10 metres above sea level.

Your PhD thesis was on the diffraction of seismic waves and you’re familiar with laws describing the dispersion of water waves — were these things rattling around in your head?

Not at those moments, no. I was, later, trying to figure out the pattern to the waves. After the big one, there were further waves at 11:10 and 11:50. These were just as big as the first wave that hit the swimming pool, so people began to panic again — but none was as big as the second wave. When I returned to England,

I started re-reading about tsunami theory, but at the time I wasn’t thinking about these things. I had gone on holiday to forget about them.

Some say that your efforts saved hundreds of lives. Is that really the case?

Well, hundreds is certainly an exaggeration, but maybe a few. It wasn’t me alone — it was also my wife and then the staff who responded so quickly. I have also read newspaper articles about a young girl who had studied tsunamis in school and told her parents what was going on; they also immediately went to a higher floor. It goes to show how useful education can be.

David Cyranoski is *Nature’s* Asian-Pacific correspondent.

Where were you when the tsunami first hit?

It was 9:30 a.m. and we — my wife Lillian and I — were just finishing breakfast in the ground-floor restaurant of our hotel, which overlooks the beach and the swimming pool. The earthquake had occurred at 7:00, but we didn’t feel it. The sea rose a few metres, although we didn’t notice until a fairly small wave rolled gently through the swimming pool and lobby of the hotel. Everyone just stopped what they were doing and watched in amazement. It only left a couple of inches of water in the lobby but there was sand and debris in the swimming pool. When we asked the hotel staff if this had ever happened before, they said “never”. People talked of a particularly high tide, but that just didn’t seem right to me.

What were your initial thoughts?

I had read about tsunamis when I was a student, especially from the big earthquake that occurred in Alaska in 1964. But I had never experienced one. I said to my wife that there must have been an earthquake in the Indian Ocean. When she asked if there was more to come, I said possibly but that it was probably from a small earthquake on



At the scene: Lillian and Chris Chapman saw the tsunami arrive.

A right to voice dissent against the establishment

Is science the new dogma? Can religion coexist with progress? You debate the issues.

Sir—Your News Feature “Studies of faith” (*Nature* 432, 666–669; 2004) is right to call science “the orthodox worldview” of the industrialized world and in many ways, it has also become, to use Tom Wolfe’s phrase, “a court from which there is no appeal”.

As you note in your Editorial, “Where theology matters” (*Nature* 432, 657; 2004), this is perhaps most clearly seen in medical research. It is often presented as being carried out purely to relieve pain and

maximize personal autonomy. Yet most religious traditions would disagree with these aims, suggesting that well-being additionally depends upon other, ‘spiritual’ factors such as expressions of love and fulfilment of purpose.

Critical dissent has played a central role in advancing scientific understanding, and the right to dissent should be held in high esteem by scientists. In the past this dissent has primarily been by thinking scientists against the religious establishment.

It seems ironic that these roles have now been reversed, with much dissent coming from thinking religious communities against the scientific establishment.

Like it or not, such dissent should be accepted, perhaps even embraced, since it may provide a means to a more balanced view of the place of science in society.

Ben MacArthur

Bone and Joint Research Group, School of Medicine, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK

Science flourishes in a secular democracy

Sir—I read with great interest the Commentary article “Time for enlightened moderation” by A. Rahman and A. Nasim (*Nature* 432, 273–274; 2004), which calls for Islamic nations to renew and reaffirm their commitment to science, in order to achieve socio-economic modernization and to combat fundamentalism, extremism and terrorism. It gives an excellent historical perspective on the relationship between science and Islam, provides accurate statistical facts and figures related to the scientific output of Islamic nations and outlines a realistic agenda for the future.

However, it misses two key elements that recent history has proven to be essential in moving forward in science: secularism and a working democracy, as exemplified by Turkey.

As the Commentary acknowledges, Turkey is the only member of the Organization of the Islamic Conference (OIC) states with universities ranking among the world’s top 500, and it leads OIC states in terms of annual output of research papers, according to Thomson ISI.

Turkey’s leading position among the OIC member states owes much to the reformer Mustafa Kemal Atatürk, who established the modern Turkish Republic in 1923. A unique feature of modern Turkey, among OIC members, is the constitutional secularism that forms the basis of the state. This is further strengthened by a genuine democracy that has its roots in the transition to a multi-party democratic system, steered by Ismet İnönü, after the Second World War.

In addition, the Turkish Republic has given the utmost importance to university education. For example, the university reform introduced by Atatürk in 1933 was a significant step forward in the pursuit of scientific excellence, and Istanbul University became a refuge for many

renowned Jewish scientists who fled Nazism in Europe. Later, the ‘university project’, led by Ihsan Dogramaci as president of the Council of Higher Education from 1981 to 1992, increased the number of Turkish universities to a level that nearly tripled the per capita figure for the OIC’s inhabitants. At the same time, the establishment of university research funds and of the Technical Research Council were instrumental in increasing Turkey’s output of scientific papers.

Turkey can be a role model for Islamic nations, striving to correct the false image of Islam as being linked with extremism, fundamentalism and terrorism.

İclal Büyükdere-Özcelik, Tayfun Özcelik
UNICEF National Committee and Bilkent University, Bilkent, Ankara 06800, Turkey

It’s not just theologians who are morally troubled

Sir—Your Editorial “Where theology matters” (*Nature* 432, 657; 2004) fails to mention that it is scientists, not theologians, who are out of step with society. The seemingly important ethical question, “Why [should society] be denied a medical advance just because some of its members find it morally troubling?”, is disingenuous.

I question the assumption that only a small minority are troubled by the ethics of medical research. In the United States, scientists who believe that “all scientifically sound lines of research should be pursued simultaneously” are in the minority. Although US polls reveal a large majority in support of stem-cell research for therapeutic purposes, they also indicate broad support for President Bush’s stance on federal funding restrictions. Scientific progress within strict ethical limitations seems to be the majority opinion.

Thankfully, we live in a democracy

where public policy is decided by elected representatives, not a scientific oligarchy. A better question is why certain individuals should be allowed to pursue a line of research when most members of our society find it morally troubling.

Stephen J. McSorley

University of Connecticut Health Center, Department of Medicine, Division of Immunology, 263 Farmington Avenue, Farmington, Connecticut 06030-1319, USA

Eastern creeds are less dogmatic about scripture

Sir—Your Editorial, “Where theology matters” (*Nature* 432, 657; 2004) is surprisingly biased towards the ‘religions of the book’ that originated in west Asia, and to Christianity in particular.

It is surprising because your News Feature “Studies of faith” (*Nature* 432, 666–669; 2004) in the same issue mentions both the Buddhist and Hindu approaches to stem-cell research, and another article, “Buddhism on the brain” (*Nature* 432, 670; 2004), describes the Dalai Lama’s interest in and approach to science.

An increasing amount of science is done in east and south Asia, and many scientists in the West (particularly the United States) are emigrants from those countries. To the extent that they are religious at all, followers of these religions (Buddhism, Hinduism and others) tend to be less dogmatic and more philosophical — less insistent on following the ‘holy writ’ of ancient texts and more in favour of searching for one’s own path in the modern world, consistent with certain basic ideas of ethics. Thus, their answers to the issues you raise in the Editorial would be quite different.

Rahul Siddharthan

Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600 113, India

Weapon of mass attraction

Scientists should embrace, not fear, television news.

Eliene Augenbraun

If the press is vital for fostering debate in a healthy democracy, then science is in trouble — because the press provides little coverage of science for the general public. I was a scientist and part-time science illustrator before moving into the media full-time in 1998. As a practising scientist, I admit I thought very little about reaching out to the public. According to a poll of members of the US scientific society, Sigma Xi, I was in good company¹. In this poll, commissioned by Research!America, 74% said they did not have time for public outreach, and 41% believed their involvement makes no difference. Half of the respondents did not know how to become involved. But after spending two years in Washington DC as a science policy fellow for the American Association for the Advancement of Science, I realized how important it is for scientists — or anyone with a strong point of view — to communicate directly and effectively with the public.

I believe that when trying to reach the general public, scientists and institutional press officers should focus on the medium that is capable of reaching the most people. Right now, this is commercial television. Scientists are needed to help this medium shed its reputation as a 'weapon of mass distraction'. This is what led me to co-found ScienCentral, a firm that focuses on short-format video for television and the Internet. Our main product is short news items for local television news in the United States and for the web; each one is seen by millions of viewers.

Scientists often look down on commercial media because they believe it lacks educational content. Many think it is better to invest resources in getting coverage in quality magazines and newspapers or specialist radio and television. Although I applaud the worthwhile projects being spearheaded by universities and science institutions — such as ResearchChannel (a cable channel in the United States, www.researchchannel.org) — and agree that it is important to provide interested audiences with in-depth content, I would argue that many more scientists need to seek breadth instead of depth.

Attention seeking

Why is squeezing a bit of science into a short television news slot between the weather and a celebrity interview worthwhile? Because whatever you think of its output, commercial television news is still a cost-effective means of educating the public about scientific find-



Show and tell: science on television reaches more people, but it needs to be visually arresting.

ings. Our stories reach millions of viewers at a cost of less than a penny per viewer. By contrast, a museum exhibit is considered successful if it attracts an audience at a cost of one dollar per visitor. In addition, television news exposes viewers to the incremental process of research — the framework on which any facts they learn must be hung.

I believe working scientists can, and should, seek out opportunities to appear on popular television as often as possible. For example, when your institute press releases

your work, it would help most scientists to talk to their press officer about the likelihood of television coverage, and what they might do to prepare for that — or even to foster it.

One of the oft-cited reasons for presenting science research to the public is that the taxpayer has the 'right to know'. In countries belonging to the Organization for Economic Co-operation and Development, taxpayers fund much of the research (60% of basic research in the United States), and are entitled to hear the results. In addition, US citizens and Europeans who are well informed about science tend to support it², whereas individuals who score low on an index of scientific optimism and high on an index of reservations tend to disagree with government support for basic research³.

Benefits of the box

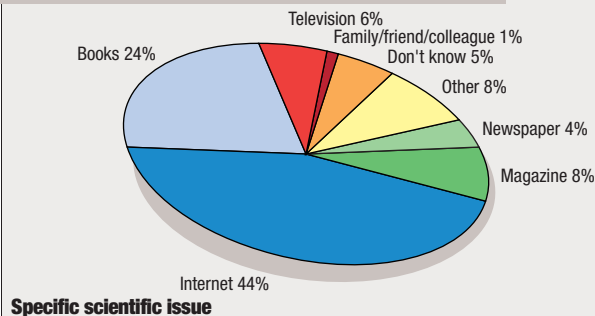
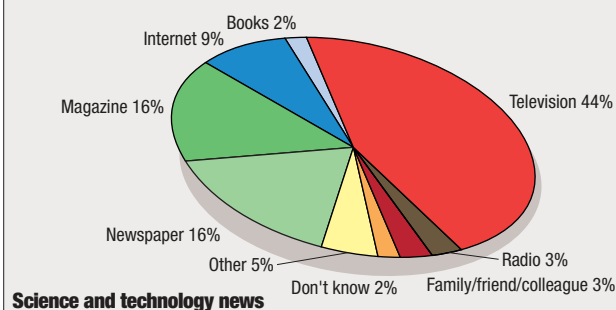
Where do most taxpayers get their news? Nearly half of the adults in the United States and Europe learn about science from television². And what they are watching, more frequently, is the lighter fare. Of course, there are highly regarded programmes available on the Public Broadcasting Service (PBS) in the United States and from public broadcasters throughout Europe. But although 8% of US adults watch the in-depth science documentary show NOVA on PBS regularly, commercial broadcast programmes (both local and network) reach far more people².

As of 2004, 59% of US adults regularly watch local television news, whereas 16% tune into network news and only 5% watch the PBS programme *NewsHour*⁴. US viewers rate local news and cable as having the highest percentage of believability out of all media³. In 2001, only 16% got their science news from newspapers, another 16% from magazines and 9% from online sources (see pie charts, overleaf)².

When a science issue is in the news, it is more likely to attract public support. For example, public access to and interest in bioterrorism news stories were high in 2001. Almost all (96%) US adults polled soon after 11 September, 2001 viewed science and technology as playing a critical role in national security². Approximately 80% felt that scientific research "is very important for meeting future terrorist threats".

A recent study of factors that increase support for science include high 'political sophistication' — an understanding of and trust in the institutions doing and reporting the work⁵. But according to Research!America, only 2% of US citizens know the names of the two main federal institutions that fund the

Sources of science and technology information in the United States, 2001



bulk of science research in the United States (the National Science Foundation and the National Institutes of Health). And we must not forget that trust in pseudoscience, such as astrology, is strong — over 40% of US citizens believe that it is 'sort of' or 'very' scientific².

Crucially, viewers remember what they see on television. When we do a story about a scientist, he or she invests hours (if not days) in preparing for, shooting and sometimes providing materials for our 90-second stories. We take 160 person-hours to complete each story. All that work for 90 seconds — does it make a difference? A recent study of Science-Central news stories by Jon Miller of Northwestern University, Chicago, determined that 4 million US adults watch any particular local news story, and 44% recall the content weeks later. These viewers do not visit science museums, botanical gardens or zoos very often — they average two visits a year to all such institutions combined (J. Miller, personal communication). So, yes, television can educate viewers — or at the very least increase their awareness of science.

If the press office agrees that your latest breakthrough could interest local television, what next? Remember, most reporters have never seen inside a science lab — their last science class was years ago. If we want the public to trust the scientific community, we must try to explain the lab and make the process of research more transparent. So invite reporters to visit, and take time to show them around.

Reaching out

Scientists and press officers need to develop relationships with journalists they trust. The journalist may not like the first, second or even tenth story you pitch, often for reasons having nothing to do with the story. Ask for feedback on why they did not run a particular idea. By inviting television reporters in and learning what they want, you will provide broadcasters with the tools for good story-telling, and ensure that they provide the public with the information they need to make informed decisions.

It also helps to understand the changing media landscape, for example, to know that news staffs are being cut. Between 1998 and

2002, the workload of each reporter increased from 1.5 to 1.8 stories a day at local television stations in the United States⁶. Science and medical news stories are often culled from wire services, but with no more research than a call to a local university, the stories are often presented without context and certainly without input from the researchers themselves.

At local news stations in the United States, reporters are very rarely assigned to specifically follow science. The closest to a science correspondent might be the health or weather reporters. Whatever they air must fit the segment of the newscast for which they are responsible. For example, climate change works nicely for weathercasters, whereas a new brain imaging technique might be useful for the health reporter. Scientists who are successful at getting on television learn to give reporters information they need, rather than what you think they should want⁷.

Making the message

Scientists and press officers can also pitch stories to specialist news organizations, such as my own (www.sciencentral.com), Discoveries and Breakthroughs Inside Science (www.aip.org/dbis), Research TV (a consortium of universities in Britain), or websites that work with television networks (www.space.com or www.discovery.com).

When you do get that first interview, there are three things to remember. First, do not forget who your audience is. Your work may be important, but that alone will not interest the average viewer. That importance must be summarized in less than ten seconds. This does not mean it needs to be simplified to the point that it is immaterial — exactly the opposite. The most important kernel of information must be teased out and presented in a straightforward manner, without jargon.

Second, make it visually attractive. No matter how important your message, the medium will not tailor itself to you. Television is fast-paced and image-based — you need to develop colourful visuals, demonstrations, analogies and anecdotes. Viewers, especially those with less education, learn from and like pictures and videos^{6,8}. For instance, the most

successful example of science education through television is the weather report. The public understands jet streams, weather fronts and barometric pressure from their graphic representations. It is your job to make it simple, because you are likely to be working with a reporter with little or no science training. A good reporter will help you to whittle down statements and pictures into small chunks, but they can only do so much.

Third, make it relevant. Most US adults (87%) want news "that is helpful in my daily life"⁶. A successful science story often answers basic questions we all ask — especially those with human interest, "Could that help my mother now? In five years?" Sometimes the idea is just really original, such as intelligent camouflage material that 'knows' into what colour it should turn itself. We call that a 'Hey Martha' story, as in, "Hey Martha, did you see that crazy thing on television?"

Local television news is popular now, but in five years, the ever changing and expanding electronic media frontier will make something else popular. More people are turning to the Internet when they want information on a specific science issue (see pie chart)². I believe that the web is great for providing background to a story, but I also see short video news reports remaining popular as they reach more wired and wireless devices. It is my hope that science will be present in this brave new world of short attention spans and instant communication. All scientists should seek the biggest audience for their work. If you don't, who will? ■

Eliene Augenbraun is president of ScienceCentral, 248 West 35th Street, New York, New York 10001, USA.

1. Brown, C. P. et al. *Sci. Commun.* **25**, 294–303 (2004).
2. NSF Science and Engineering Indicators 2004; available at www.nsf.gov/sbe/srs/seind04/start.htm
3. Miller, J. D., Pardo, R. & Niwa, F. *Public Perceptions of Science and Technology: A Comparative Study of the European Union, the United States, Japan, and Canada* (Fundacion Banco Bilbao Vizcaya, Bilbao, 1997).
4. *Biennial Media Consumption & Believability Study*; available at <http://people-press.org>
5. Sturgis, P. & Allum, N. *Public Understand. Sci.* **13**, 55–74 (2004).
6. *The State of the News Media 2004*; available at www.stateofthenewsmedia.org
7. Borchelt, R. in *Communicating the Future: Best Practices for Communication of Science and Technology to the Public* 61–68 (National Institute of Standards and Technology, Gaithersburg, 2002).
8. Sternberg, R. J. *Wisdom, Intelligence, and Creativity Synthesized* (Cambridge Univ. Press, Cambridge, 2003).

Armed only with knowledge

Physicist Patrick Blackett made a stand against atomic weapons.

Blackett: Physics, War, and Politics in the Twentieth Century

by Mary Jo Nye

Harvard University Press: 2004. 272 pp.

\$39.95, £25.95, €36.90

Jon Agar

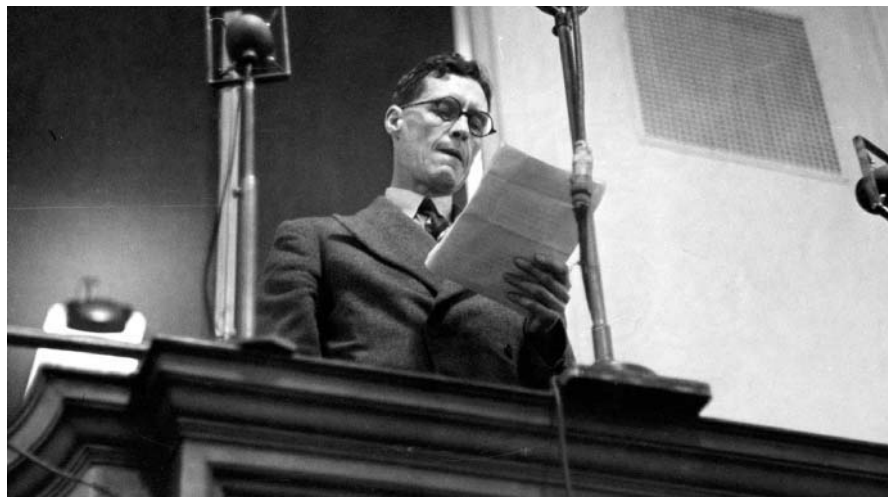
A prominent scientist visits the United States at election time and records a “conservative turn in American politics and an eagerness to talk about preventive war”. Rather than trim his sails to meet the prevailing wind, he redoubles his determination to speak the truth about weapons of mass destruction. He subsequently writes a book that is criticized from all sides, and is labelled a defeatist, even a traitor of sorts. It will take many years for the historical consensus to catch up.

The scientist was Patrick Blackett, and the book was his 1948 work *The Military and Political Consequences of Atomic Energy*. “The experiences of Blackett in his public campaign against nuclear weapons,” writes his biographer Mary Jo Nye, “illustrate the risks to a physicist of writing about a subject other than physics, as well as the circumstances that might compel one to do so.”

The traditional function of a biography was to provide lessons, drawn from a past life, by which contemporary readers should aspire to live their lives. The emphasis was on character. Nye’s biography of Blackett, although modern in its approach to the history of science, is nevertheless oddly traditional. In seeking to provide an account of how the physicist could navigate so successfully between theory and experimental craft, and between the laboratory and the corridors of power, she focuses on his virtuous characteristics, including “versatility of imagination”, “tough skepticism” and being “fiercely independent”.

The children of a century ago were blessed: the most radical technologies of their day could be taken apart, built, rebuilt and understood with their own hands. “I spent every hour out of school making wireless sets and model aeroplanes,” Blackett recalled. The knowledge gained from this tinkering got him into the Royal Naval College at Osborne House on the Isle of Wight: the admissions board judged him the “right sort of boy”. The young naval cadet built his own cameras. On active duty he witnessed the Battle of the Falklands in 1914. In quieter moments he continued with photography.

Blackett’s manipulative skill flourished when, after demobilization, he joined Ernest Rutherford’s Cavendish Laboratory at the University of Cambridge, where the most



Speaking out: Patrick Blackett was labelled a traitor for arguing against UK work on an atomic bomb.

remarkable series of discoveries in twentieth-century physics were crafted.

Blackett constructed automatic cloud chambers in which the passage of a particle triggered a photograph of the event. Working with Giuseppe Occhialini, he turned the device towards the mysterious cosmic rays. “What not everyone had the chance to see,” Occhialini later noted, “was the passionate intensity with which he worked. I can still see him, that Saturday morning when we first ran the chamber, bursting out of the dark-room with four dripping photographic plates held high, and shouting for all the Cavendish to hear, ‘One on each Beppe, one on each!’”

But, as Nye writes, intensity was balanced by tough skepticism. In 1932, faced with a small number of extraordinary tracks, Blackett identified them with one of his colleague Paul Dirac’s most exotic predictions, positively charged electrons. Yet he refused to rush into print — one reason why the 1936 Nobel prize went to his rival at the California Institute of Technology, Carl Anderson.

Francis Everitt, one of Blackett’s students, claimed that two remarks encapsulated Blackett’s outlook: “Make sure you gather plenty of data,” and “Treat your research like a military campaign.” Others, too, clearly learned the lesson. When Blackett restarted his cosmic-ray programme at Manchester University after the Second World War, George Rochester and Clifford Butler soon announced their discovery of kaons, or K-mesons, in *Nature*: “There are two photographs containing forked tracks of a very striking character. These photographs have been selected from five thousand photographs taken in an effective operation time of 1,500 hours”. Plenty of data there.

Nor is it surprising that the ex-naval cadet treated research like a military campaign. However, during the war Blackett achieved something with far greater ramifications: he treated military campaigns as research, inventing a new science, operational research, as a result. In 1933, Blackett had moved to Birkbeck College, London. The problems he faced of limited financial support for his impoverished physics laboratory were easily outweighed by the advantages of his “fierce independence”, free from Rutherford and free to choose his own research direction.

London also meant induction into the political establishment. The socialist Blackett was recruited by Henry Tizard to advise on air defence, in particular the development of radar. After war broke out, he assembled interdisciplinary teams of scientists to work closely with military personnel. The stuff of the operations rooms — flight books, radio-location data and ship positions — became the data of operational research. This was a new science, not because its approach was new, but because it marked a close and supplementary relationship between scientific expertise and military power.

When Blackett wrote in November 1941 to his friend Michael Polanyi, an émigré chemist of opposing political views, Blackett asked: “You speak as if it is always a duty to publish the ‘truth’. If I had published the truth of what I have known of parts of our war effort, I would certainly be locked up. Should I have done so?”. What was the truth in question? Blackett had sided with Tizard in opposing Frederick Lindemann’s advocacy of bombing civilian populations, and he had done so on the basis of operational research. By then, other means of mass destruction were being debated.

From 1940, Blackett sat on the MAUD Committee, assessing the likelihood that research on nuclear chain reactions would lead to a practical atomic weapon within the timespan of the war. At first he was the lone British voice calling for the weapon to be developed only by the United States. After the war, and particularly after the 1946 McMahon Act broke with any pretence of joint UK-US responsibility for the bomb, Blackett argued vehemently against a British bomb project. He tried private routes of influence, but was rebuffed by the prime minister, Clement Attlee. So he went public, writing *The Military and Political Consequences of Atomic Energy*.

"Neither communist nor pacifist, Blackett had no argument with war," writes Nye, so why did Blackett take such "an outspoken and unpopular political position on matters of nuclear policy immediately following the Second World War?" Because his naval and operational-research experience taught him that policy decisions driven by inadequate knowledge were likely to be wrong. And because he was appalled by war-games theorizing, which he viewed as inhuman. ■

Jon Agar is in the Department of History and Philosophy of Science, University of Cambridge, Cambridge CB2 3RH, UK.

Living with viruses

Viral Fitness: The Next SARS and West Nile in the Making

by Jaap Goudsmit

Oxford University Press: 2004. 202 pp.

£18.50, \$29.95

Steven Wolinsky

Breathing can kill you. So can eating and drinking. We live in a world where pathogenic microorganisms in air, food and water pose an omnipresent threat to human health and agriculture. Yet we continue to expand our global presence, engage in high-risk sexual behaviour, and produce more crops and domesticated animals bred for traits that restrict their diversity. As a result, we are exposing ourselves to dangerous viral pathogens that can cause epidemics on a scale seen only in apocalyptic novels. Viruses will inevitably help decimate our natural world and humans as well. So claims Jaap Goudsmit in his engaging new book, *Viral Fitness*.

Goudsmit, a professor of communicable diseases at the University of Amsterdam in the Netherlands, chooses several diseases of plants, animals and humans as case studies in the epidemiology and evolutionary biology of emerging viral pathogens. He highlights important ecological factors in the emergence of viruses, such as the role of waterfowl in the rise of the H5N1 influenza virus,



Paying the price: the H5N1 influenza virus that caused avian flu was spread in Chinese markets.

which led to an outbreak of avian flu in Asia; the part the bushmeat trade played in the appearance of HIV in Africa; and the role of the consumption of palm civet in the spread of the coronavirus that causes severe acute respiratory syndrome (SARS).

To become dangerous to humans, such viruses must cross the species barrier, which requires both genetic factors and incursions into new ecological niches. Antibodies against the SARS coronavirus from the Himalayan palm civet, the putative species of origin, have been found in other animals sold at local Chinese markets, implying that there are constraints to adaptation across species. Goudsmit points out that few transmission events are sustained. Monkeypox virus and the O and N groups of HIV type 1, for example, infected human hosts with limited subsequent transmission. This suggests that dead-end transfers of viruses with imperfect adaptation to the new host species may be common, and that transmission of a pathogen that spreads to epidemic proportions may be the exception rather than the rule.

Goudsmit deftly reconstructs epidemiological history and relates how climatic changes, population movements and trade have converged to help viruses emerge and spread. Many of these anecdotes are well known, but others are not. Goudsmit artfully unravels the threads that tie together the evolutionary selection processes working in the new host species. Because viruses have large population sizes, high mutation rates and short generation times, they are capable of rapid genetic evolution. Once inside the host, virus populations are shaped by forces of evolutionary change that include mutation, genetic recombination and natural selection. This complex interplay between the virus and its host — both in a single individual and in the population — can result in a variety of outcomes. For example, the introduction into Australia of a myxomavirus to reduce the rabbit population was highly

successful, through an accidental experiment of nature. At first, rabbit numbers were drastically reduced. Over time, however, a milder strain emerged that was more effective at infecting rabbits. Through selection, the virus evolved to a less virulent form, illustrating the important difference between evolutionary fitness and virulence.

Despite the fascinating examples he cites, Goudsmit fails to address some critical topics, such as the contribution of host and virus genetic heterogeneity and coevolution, and the role of frequency-dependent selection in evolutionary change. Several of his suggestions are untenable, such as the idea that a new virus can emerge after an asexual *ménage à trois* among unrelated viruses in a single cell; not every virus can infect every cell. Viruses have anthropomorphic desires and a teleological end in view, according to Goudsmit. Other topics, such as the role of viruses in making possible our evolutionary development, and the use of phage therapy for clinical and agricultural applications, add another dimension to the host-pathogen relationship.

To bolster the claim that viruses are a threat to us now "more than ever before", Goudsmit considers epidemiological and evolutionary dynamics alongside the course of human events, but neglects to mention many public-health successes. Health officials are scrambling, so far with relative success, to contain the SARS coronavirus and prevent the spread of the influenza H5N1 and H7N7 viruses from waterfowl to humans. No mention is made of the important change to seasonal outbreaks of influenza achieved by simply moving pigs away from ducks on Chinese farms.

How can we avoid the dangers that nature presents? Wash your hands. Cover your mouth when you sneeze. Refrain from transplanting animal organs, Goudsmit would also add, and don't eat monkeys. Vaccines help to halt viruses that cause epidemics such as measles, which cause short infections

with strong cross-immunity, and influenza. For retroviruses such as HIV that lead to persistent infection, the prospects for a vaccine are dim; their diversity exists both in the individual and in the population as a whole. Until we have an effective AIDS vaccine, people will need the education and resources to modify their behaviour. Unless we change our way of life, Goudsmit warns, the emergence of viral threats to human health looms large. ■

Steven Wolinsky is in the Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA.

Peering out of the box

Pioneering Research: A Risk Worth Taking

by Donald W. Braben

Wiley: 2004. 198 pp. £23.50, \$39.95, €33.30

Nina Fedoroff

Donald Braben's book begins with the idea that the unique quality of *Homo sapiens* is not wisdom, but rather the capacity for dissent. By this he means not the transient dissatisfactions that spark violence at sports events, for instance, but the dissent that emerges from "an individual's overwhelming conviction that some aspect of life has become unbearable". Braben believes it is this characteristic of humans that gives rise to scientific discovery of the kind that changes our view of the world, be it Galileo's restructuring of the Universe or Barbara McClintock's transposons, which overthrew the notion that genomes are static. And because it is increasingly recognized that science and technology power contemporary economic growth, it follows that the freedom to challenge the prevailing thinking is essential not just to deepen our understanding of the world, but to sustain economic growth as well.

Suppressing dissent leads to both scientific sterility and economic stagnation, be it in Europe in the Dark Ages or in the Soviet Union not so very long ago. But, more ominously, Braben argues that what suppresses scientific invention today isn't religious dogma or a political regime (although some contemporary observers of US politics might argue otherwise). Rather, it is the well intentioned bureaucracy of the system for awarding grants and the way it operates in a resource-limited environment that is at once egalitarian and increasingly mindful of cost-effectiveness.

Braben argues that until the last few decades, scientists with unconventional ideas could afford to ignore the opinions of their colleagues. But not any more. This is

Video installation

Twin peaks

London Fieldworks artists Bruce Gilchrist and Joelson have explored the work of two scientists who studied the weather from mountain-top observatories in the nineteenth century, and who went on to develop instruments that presaged the development of particle physics and space plasma physics.

C. T. R. Wilson observed visual phenomena such as the 'Brocken spectre' in the skies above Ben Nevis in Scotland when working as a relief meteorologist. He went on to develop the cloud chamber, which enables the visualization of the tracks of subatomic particles, earning a Nobel prize for his efforts.

Kristian Birkeland's observations of the aurora borealis from the summit of Haldde in Norway inspired him to build the terrella,

which models the aurora's relationship to solar activity.

Little Earth, a video installation that explores this move from lone observers of nature to an era of technological and abstract science, can be seen at the Wapping Project in London until 12 February, and then at the Fort William Mountain Film Festival in Scotland until 3 March.

♦ www.londonfieldworks.com



because "selected groups of their fellow experts now have responsibility for setting priorities and allocating resources. Consensus opinion cannot be ignored." What he's talking about, of course, is the peer-review system: panels of experts who gather all over the planet to decide which research proposals are worth funding and which are not. But, Braben says: "An expert opinion is one thing; the consensus of experts is another." He goes on: "Unfortunately, science and democracy are poor bedfellows... No matter how many agree on the validity of a point of view, a single person with a more viable, accurate, or comprehensive alternative may overthrow it." Well said, and all too true.

One might question Braben's anthropology and history, but he certainly has the credentials to comment on contemporary scientific bureaucratization. After almost two decades as an academic nuclear and elementary particle physicist, he joined the ranks of scientific administrators, first in the UK Cabinet Office and later in the Science Research Council. There he had responsibility for the Marine Technology Directorate, which funded marine technology research, and the Teaching Company, whose objective was to bring academic engineers together with industry. He also worked as chief scientist at the Bank of England printing works.

Braben was then recruited by the energy company BP in 1980 to head its 'blue skies' research initiative, which came to be called the Venture Research Unit. The unit's mission was simply "to support the research that might lead to new types of industrial activity". Constrained by neither a specific mission nor an 80-page manual for the preparation of grant proposals, but only by a limited budget, he could make it up as he went along. Well, almost: he still had to convince the members of BP's Venture Research Advisory Council, all of whom were fellows of the Royal Society, and all of whom objected to

his abolition of peer review as the mechanism for funding research, a system they didn't think needed fixing.

But instead of focusing on grant proposals, Braben insisted on choosing people, treating their selection "as if it were a scientific problem rather than an administrative one" and struggling "constantly to reduce the number of rules" imposed on applicants. In Braben's view, "the highest accolade one can give a new research proposal is that it could radically change the way we think about something important." That's why he sought researchers whose thinking was decidedly out of the box, working diligently to make sure that it wasn't merely out to lunch — not an easy task on which to score well.

I unreservedly recommend this book to anyone who has puzzled over the growing malaise of contemporary scientific research: we build virtual mountains of data, but where are the paradigm-busters? Braben says they're still there, but that the system is almost guaranteed to filter them out.

I think he has got the diagnosis right. Does he also have the cure? I leave it to others to judge Braben's personal success rate, but the prescription is certainly inviting: pick people and their ideas, not projects; trust them, they'll know when it's time to stop or change direction; provide freedom and sufficient money; expect radical ideas, and expect them to meet with resistance, at least initially. That seems a bit revolutionary to those of us who have grown accustomed to the proposal grind: specific objectives, preliminary data, experimental plan, deliverables and timeline. Small wonder then that there's no time to wander off the beaten path.

Oh, and did I say that this book is a surprisingly good read? Braben is literate, pithy and personable. ■

Nina Fedoroff is at the Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

Schrödinger's mousetrap

Part 2: Enter the detective.

Neil Mathur

Wearing his long brown coat and crumpled features, Karl Lister looked out from the platform at the sea of faces. He had never seen so many physicists and they were not like he imagined. Indeed, only one of them had a beard ... but he was lying dead on the floor with a hole in his head, made by the laser. And although the burning odour had now gone, the impenetrable detective smelt murder.

While his colleagues interviewed members of the audience, a study of the video footage allowed Lister to establish that a rather sophisticated crime had taken place. If Rufus Jaeger was indeed the intended victim, the overly powerful laser was set up with a knowledge of the seating plan for the platform. After all, he reasoned, one would hardly engineer the whole scenario just to make a random hit. The more he learnt about scientists, the more they frightened him. No wonder someone had once called physics the science of death.

Lister decided to begin his questioning with Jaeger's head technician, Tony Trotman, as he was the person responsible for the fatal demonstration. Trotman's silver hair and craggy features exuded a strange serenity that Lister registered at once.

"Mind if I smoke?" asked Trotman.

Lister flinched but said he didn't mind, as he wanted Trotman to relax fully and open up to him. The nicotine, however, immediately produced the opposite effect. Before Lister could ask his first question, Trotman began to look agitated and started muttering something that finished with "... and I need this like I need a hole in the head". Immediately regretting this choice of words, he by turns reddened, coughed and pretended to have something in his eye.

Impassively, Lister opened the interrogation by asking whether Trotman had been close to the deceased. Becoming animated, Trotman launched forth.

"Not exactly. To be honest, maybe I shouldn't say this, but nobody liked Rufus very much. Except for his postdoc Ludmilla, but she likes everybody."

A frown spread across Lister's face. "What's a postdoc?" he asked.

"It's basically someone in the lab who doesn't have a cat in hell's chance of getting

a proper job." Lister made a mental note that this reference to cats and probability could be significant.

Refocusing, Lister asked Trotman to describe the fatal demonstration from his perspective. Trotman explained that he had taken his cue from the end of the multimedia presentation.

"And then what?" asked Lister.

"And then ... I proceeded to reset everything for a second trial..." Trotman faltered, "since a second trial had been agreed."

"You didn't notice anything untoward after the first one?"

"Certainly not," replied Trotman with total confidence.

Lister proceeded to ask Trotman a series of questions about the set-up. Yes, it was Trotman who supervised its construction, but Jaeger had the final say. No, it was an external supplier who provided the prisms. "And who decided which laser to use?" asked Lister.

"Funny you should ask," said Trotman darkly. "The laser that Rufus planned to use was a much more modest affair, but it stopped working two days ago. So we ended up using our other one which is ... much more powerful." Regaining his composure, he went on. "Look, Rufus put his faith in me. We never had any accidents. Ever. I feel awful about what happened. But I can't see what went wrong. With half an hour to go, Rufus and I tested the set-up and everything worked fine."

Lister wanted to know what happened

next. "The two of us left the room together. As far as I know, the hall was then empty till the start of the lecture." Where did Jaeger and Trotman go? "Rufus went to join the coffee break in the hallway next to the lecture hall. As for me, mingling with scientists is not my cup of tea so I went for a smoke outside." Had Trotman seen anything suspicious? No ... well, he did see Jaeger's next-in-line Wilfred de Bruijn barging away from the coffee break angrily. "Then again, Wilf always seems to be angry about something."

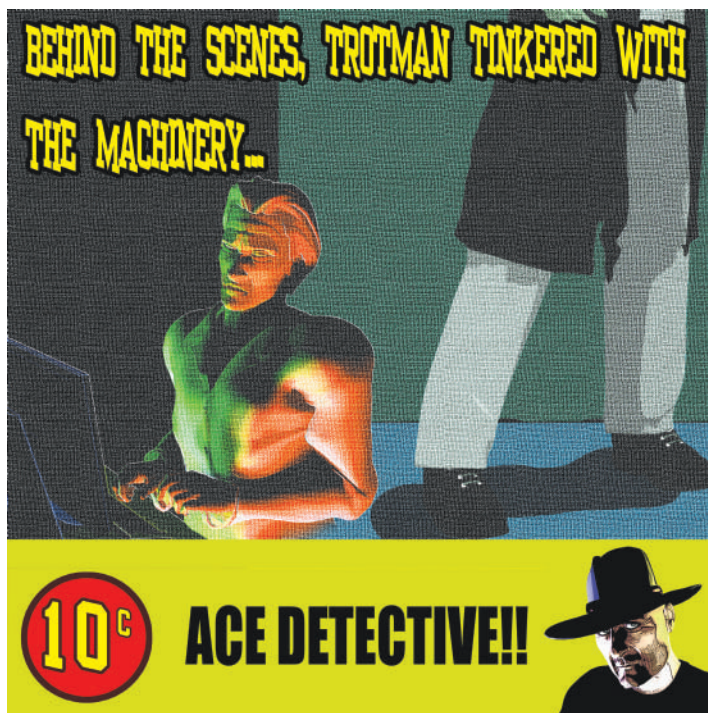
Experts arrived on the scene and with Trotman's supervision they carefully set about testing the 'mousetrap'. Lister meanwhile was making notes. If Trotman was telling the truth, there was about half an hour during the coffee

break to tamper with the set-up, assuming that it was not already compromised. But in what way had it ever been compromised? Trotman and the experts found that although the laser still pointed in the wrong direction, all the prisms seemed mysteriously to be exactly where and how they should be. Things were looking complicated and Lister's mind came back to the many factors that had come together to ensure the murderous outcome: the powerful laser, the seating plan for the platform, the multiple tests ... now hadn't Jaeger said something in his presentation about superposition effects?

Lister wanted to know how many times Trotman had tested the set-up. The answer was twice. Then, with an insight that betrayed his grasp of quantum mechanics, Lister threw another question at the technician: "And after those two tests was the mouse 'dead' or 'alive'?" Trotman replied that the first test had given 'alive', with the amusing consequence that Jaeger had declared the need to increase the budget for cheese on future grant applications. "And what about the second test?" demanded Lister impatiently. Sounding perhaps slightly less self-assured than before, Trotman explained that he had been hidden behind the apparatus and couldn't actually see the hologram projections himself.

To be continued...

Neil Mathur is in the Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB2 3QZ, UK.



Romanesque networks

Steven H. Strogatz

Newly won evidence shows that many real-world network systems obey a power-law scaling, just as if they were fractal shapes. Could this be the harbinger of a new architectural law for complex systems?

At lunch a few weeks ago, a colleague abruptly pulled a massive vegetable from his briefcase and passed it around the table. He was offering the rest of us a head of Romanesque broccoli — to ponder, not to eat.

The surface of a Romanesque is exquisitely symmetrical, sporting dozens of knobbly florets, each a miniature version of the entire structure, and each built from even smaller copies of the whole (Fig. 1). On page 392 of this issue, Song, Havlin and Makse¹ show that this same kind of symmetry, called self-similarity, is shared by a wide array of networked systems, from the hyperlinked pages of the World Wide Web to the biochemical reactions underlying cellular metabolism.

The pervasiveness of this symmetry hints at a new architectural law for complex systems. Roughly speaking, many large, interconnected systems are tied together in the same way across increasing levels in their hierarchical organization. The links between clusters of nodes, and between clusters of clusters, and so on, obey the same statistical trends as the links between individual nodes themselves.

To quantify these ideas, Song *et al.*¹ borrow tools from fractal geometry and statistical physics. Fractals are self-similar shapes². They have been studied for decades, first in pure mathematics (where they were initially derided as monsters), and later in the natural sciences, to help researchers analyse such erratic phenomena as the 'burstiness' of Internet traffic and the shapes of cities as they grow.

A fractal geometer might characterize the roughness of Romanesque broccoli by computing its 'box dimension', as follows. Represent the broccoli's surface as an enormous collection of points, and divide the space around the surface into a fine lattice of cubic boxes, a three-dimensional analogue of graph paper. The boxes that intersect the broccoli's surface contain points; the others remain empty. The key is that if you make the boxes bigger, you need fewer of them to cover the same set of points. Specifically, the number of occupied boxes N decreases with ℓ , the side length of the cubes, according to a power law: $N(\ell) \propto \ell^{-d}$. The exponent d is the box dimension. For classical, non-fractal shapes, d reduces to the usual dimension: $d=1$ for a line or a smooth curve, and $d=2$ for a



Figure 1 Fractal vegetable. Fractals are intricately repeated shapes, like the surface of this Romanesque broccoli, in which the parts resemble the whole across several levels of resolution. The work of Song *et al.*¹ indicates that many complex networks, from protein–protein interaction maps to the collaboration graph of Hollywood film actors, are 'self-similar' in much the same way.

plane or a smooth surface. But for the wildly corrugated surface of the Romanesque, it turns out that $2 < d < 3$, a result that strikes the uninitiated as bizarre.

To generalize this box-counting method to networks, Song *et al.*¹ had to define a suitable collection of non-overlapping boxes of size ℓ . The difficulty is that for most social or biological networks, there is no natural euclidean space surrounding the network, so ℓ has to be defined intrinsically, using the network itself. The solution is to define the distance between two nodes as the number of links in the shortest chain between them. For instance, in a friendship network your immediate friends are one step away from you, friends of a friend are two steps away, and so on. Next, Song *et al.* enumerated all the ways to partition the network into distinct boxes of size ℓ . The rules are that two nodes can be in the same box if the distance between them is less than ℓ ; each node must be in some box; and no node is allowed to be in two or more boxes.

For a large network, there are a tremendous number of such partitions. By exhaustive searching, Song *et al.* found the partition requiring the fewest boxes, and then examined how this minimal number N depends

on the resolution ℓ . They found that many (but not all) real-world networks obey a power-law scaling, $N(\ell) \propto \ell^{-d}$, just as if they were fractal shapes.

More evidence of self-similarity came when Song *et al.* 'renormalized' their networks. Adopting a standard technique in condensed-matter physics, the authors now ignored the microscopic details of how the nodes were connected within a given box, and focused instead on how the boxes were connected to each other. (Two boxes are defined as being connected if any of their constituent nodes are linked.) In effect, renormalizing the network coarsens it, with boxes now playing the role of super-nodes. The coarsened networks satisfy the same power-law scaling, with the same value of d , suggesting that boxes are wired together in the same way as individual nodes.

These results raise some puzzles of interpretation. First, what do the boxes really mean? Although it's tempting to think of them as something like modules³, communities⁴ or network motifs⁵, the boxes have no known functional significance, raising the possibility that the observed self-similarity might contain an element of numerology. On the other hand, some real networks

(for example, the Internet backbone) and some standard models (such as random graphs⁶ and preferential attachment models of scale-free networks⁷) lack the self-similarity reported by Song *et al.*, indicating that such scaling is not automatic and, consequently, that it can be used as a benchmark for testing models of network structure. Second, it's odd that networks should find themselves configured as fractals. In statistical physics, power laws and self-similarity are associated with phase transitions — with systems teetering on the brink between order and chaos. Why do so many of nature's networks live on a razor's edge? Have they self-organized to reach this critical state⁸, perhaps to optimize some aspect of their performance, or have they merely followed one of the

manifold paths to power-law scaling⁹, full of sound and fury, signifying nothing? ■

Steven H. Strogatz is at the Center for Applied Mathematics and the Department of Theoretical and Applied Mechanics, 212 Kimball Hall, Cornell University, Ithaca, New York 14853-1503, USA. e-mail: strogatz@cornell.edu

1. Song, C., Havlin, S. & Makse, H. A. *Nature* **433**, 392–395 (2005).
2. Falconer, K. *Fractal Geometry: Mathematical Foundations and Applications* (Wiley, Chichester, 1990).
3. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. *Nature* **402**, C47–C52 (1999).
4. Girvan, M. & Newman, M. E. J. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
5. Milo, R. *et al.* *Science* **298**, 824–827 (2002).
6. Erdős, P. & Rényi, A. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
7. Barabási, A.-L. & Albert, R. *Science* **286**, 509–512 (1999).
8. Bak, P., Tang, C. & Wiesenfeld, K. *Phys. Rev. Lett.* **59**, 381–384 (1987).
9. Newman, M. *Nature* **405**, 412–413 (2000).

Sexually transmitted diseases

Epidemic cycling and immunity

Bryan Grenfell and Ottar Bjørnstad

Are syphilis epidemics caused by external factors such as human sexual behaviour, or are factors intrinsic to the pathogen more important? Comparing the dynamics of syphilis and gonorrhoea provides some clues.

The great Renaissance physician and scholar Girolamo Fracastoro achieved lasting fame for his early observations on the contagion theory of the transmission of infectious disease¹. In 1530, he also coined the name for syphilis — which was then spreading rapidly through Europe — in an extended allegory written in Latin hexameter (epidemiologists were more culturally rounded in those days). In Fracastoro's poem, the god Apollo is angered by a shepherd, Syphilus, and afflicts him with the new disease. On page 417 of this issue, Grassly

*et al.*² provide a more down-to-earth explanation of the dynamic processes underlying the incidence of syphilis.

Whether fluctuations in epidemics are governed by external drivers (such as behaviour, climate — or the gods), or by intrinsic processes that arise from the dynamic feedback between host and pathogen populations, has been debated since the early 1900s. This controversy parallels long-standing ecological arguments about the relative role of extrinsic (environmental) forces and intrinsic, nonlinear dynamics in driving

population fluctuations³. Arguably, ecologists are more familiar than epidemiologists with the potential of nonlinear dynamics to drive cycles. This is ironic, because many infectious diseases have excellent historical records of incidence and a simple natural history — an ideal combination for exploring the underpinnings of dynamic fluctuations^{4,5}.

Comparative approaches, where differences in the dynamics of various infections can be related to biological differences in the underlying host–pathogen interactions, are particularly powerful in studying this problem⁶. Grassly *et al.*² use this approach to explore the dynamics of syphilis and gonorrhoea. They base their analysis on disease notification statistics from the United States, where these two sexually transmitted diseases are endemic. The authors use time series of annual disease reports for 68 US cities to demonstrate marked 8–11-year cycles in syphilis incidence from the 1960s to the 1980s. These cycles had previously been attributed to changes in factors extrinsic to the host–pathogen interaction, particularly to changes in human sexual behaviour. If this were the case, however, there should be correlated fluctuations in gonorrhoea because of its similar transmission route and infectious period. Grassly *et al.* demonstrate that there is no such correlation: gonorrhoea shows slow trends rather than cycles during the same period.

The authors use mathematical models to reveal that the distinct behaviours of syphilis and gonorrhoea arise from their different interactions with the human immune system. Thus, the simplest explanation for the periodicity in syphilis incidence is that it results from nonlinear interactions that are fundamental to the host–pathogen transmission process.

Syphilis stimulates significant — albeit

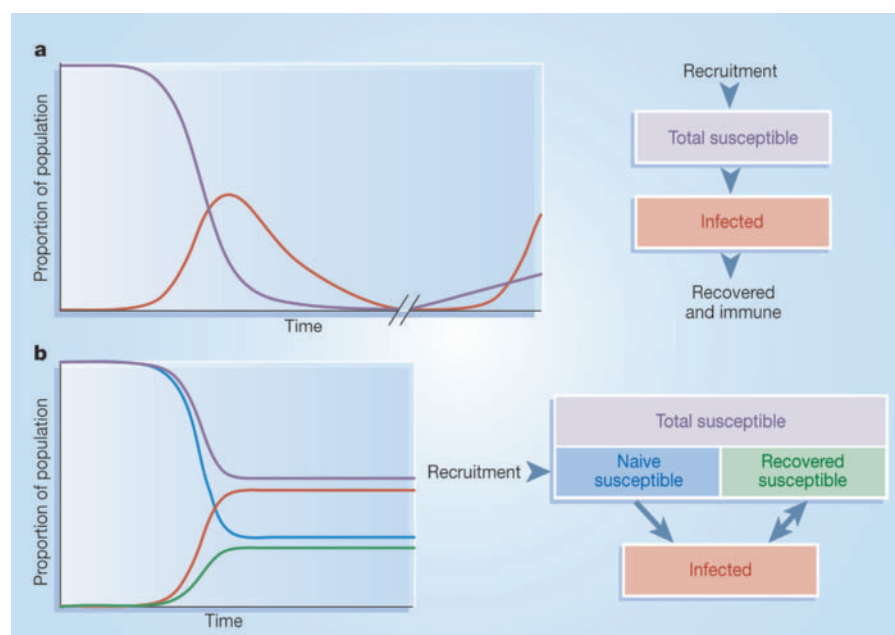


Figure 1 The impact of immunity on the dynamics of epidemics. **a**, Dynamics of the basic susceptible–infected–recovered (SIR) model, assuming that the infection attacks a naive susceptible population. Because of prolonged immunity following recovery, the supply of susceptible individuals becomes exhausted and the epidemic extinguishes itself. After the epidemic, new recruits augment the susceptible class until a further epidemic is possible (shown schematically here). This is the basis of syphilis dynamics, as shown by Grassly *et al.*¹. **b**, The susceptible–infected–susceptible (SIS) model, where there is no immunity to reinfection (as with gonorrhoea). Unlike the SIR model, total susceptible numbers are replenished by a flow of previously infected individuals, so that the epidemic moves smoothly to an equilibrium level (analogous to the carrying capacity of logistic models in ecology⁸), rather than declining.

imperfect — immunity following recovery from infection. Consequently, as Grassly *et al.* show, the dynamics of syphilis infection have many features of the well-known 'susceptible–infected–recovered' (SIR) model for microparasitic infections^{4,5}. In SIR dynamics, oscillations in disease incidence can be driven by prolonged immunity following infection (combined with a relatively short infection period⁴). Cycles occur because major epidemics extinguish themselves by exhausting their supply of susceptible individuals (Fig. 1a); the numbers of individuals in at-risk groups then build up slowly, eventually providing enough scope for the next major outbreak.

Unlike syphilis, gonorrhoea can evade post-infection immunity by camouflaging itself with different arrays of surface proteins⁷. At the population level, this corresponds to the 'susceptible–infected–susceptible' (SIS) model of infection, in which the same individual can be infected repeatedly^{4,8} (Fig. 1b). Thus in SIS dynamics, infection does not decrease the total number of susceptible individuals, preventing the boom-and-bust dynamics seen in acute SIR infections.

By contrast, the prolonged immunity seen in SIR systems causes overcompensatory dynamics and recurrent epidemics, which bear strong analogies to the cycles of many predator–prey systems in ecology. In fact, SIR dynamics are, like the SIS interaction, regulated by an upper population limit on cases and susceptible individuals⁴; the system therefore needs some form of regular or stochastic 'forcing' to drive strong epidemics. A dramatic illustration is given by acute, immunizing childhood infections such as measles, where seasonal variation in contact rates can produce violent biennial epidemics⁵. With its more sedate decadal dynamics, syphilis is unperturbed by seasonal influences. However, Grassly *et al.* show that random 'shocks' provided by demographic stochasticity (arising from the probabilistic nature of individual infection events, for example) are sufficient to excite oscillations in the model that closely resemble the cycles seen in the incidence data. On a longer time-scale, external shocks, such as a reduction in sexually transmitted disease associated with control measures against the spread of HIV, further influence the dynamics of both syphilis and gonorrhoea.

Nonlinear overcompensatory interactions, as seen in syphilis, can lead to the emergence of space–time dynamics, such as synchronization of disease incidence across different locations, or waves of infection across geographical areas. Indeed, Grassly *et al.* document increasing synchronization of syphilis epidemics across US cities during the 1960s and 1970s. The authors argue convincingly that this is because the underlying network of sexual contacts is

becoming increasingly interconnected. However, previous comparative analyses of the spatio-temporal dynamics of measles and whooping cough show that dynamic inference from such systems is not straightforward⁶. In these cases, spatial dynamics emerge through the interaction between local dynamics and spatial coupling between different local systems — sometimes coupling leads to enhanced synchrony, but sometimes synchrony can decay with time if increases in coupling accompany changes in local dynamics.

A challenge in the spatio-temporal dynamics of syphilis is to combine models for local transmission with models of spread across spatial networks⁹. As Grassly *et al.* point out, we must be very cautious in applying simple models to explain the spatial dynamics of human pathogens: basic distance-based networks fail to capture the complex nature of human population mixing. Such contact patterns are often hard to measure directly, and an intriguing task for future models will be to infer spatial patterns and temporal trends of mixing from the analysis of epidemic synchronicity.

Grassly and colleagues' dissection of cyclic versus non-cyclic behaviour is a valuable

addition to the taxonomy of comparative disease dynamics. Using comparative studies to tease out the relative role of intrinsic dynamics and extrinsic shocks is an important process for understanding and predicting the dynamics and evolution of established and emerging infections. In the nonlinear, behaviourally and environmentally driven world of epidemics, though, we should always expect the unexpected. ■

Bryan Grenfell is in the Department of Biology, 208 Mueller Laboratory, and Ottar Bjørnstad is in the Departments of Entomology and Biology, 501 ASI Building, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: grenfell@psu.edu

1. Gould, S. J. *Nat. Hist.* **109**, 38–47 (2000).
2. Grassly, N. C., Fraser, C. & Garnett, G. P. *Nature* **433**, 417–421 (2005).
3. Bjørnstad, O. N. & Grenfell, B. T. *Science* **293**, 638–643 (2001).
4. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1992).
5. Grenfell, B. T., Bjørnstad, O. N. & Kappey, J. *Nature* **414**, 716–723 (2001).
6. Rohani, P., Earn, D. J. D. & Grenfell, B. T. *Science* **286**, 968–971 (1999).
7. Snyder, L. A. S., Butcher, S. A. & Saunders, N. J. *Microbiology* **147**, 2321–2332 (2001).
8. Hethcote, H. W., Yorke, J. A. & Nold, A. *Math. Biosci.* **58**, 93–109 (1982).
9. Xia, Y. C., Bjørnstad, O. N. & Grenfell, B. T. *Am. Nat.* **164**, 267–281 (2004).

Cell biology

Border crossing

James U. Bowie

The 'translocon' complex, which determines whether a protein segment will be inserted into or pushed through the cell membrane, seems to make the decision by performing a thermodynamic measurement.

Cells have a border security system that would put any nation to shame. They need to decide millions of times every nanosecond whether to allow something in or out, and a mistake could mean death. The sentries entrusted with these life-or-death decisions are specialized proteins that reside in the cell membrane. Not just any 'wannabe' gets to be a membrane protein, however; it requires a special constitution and careful nurturing. A paper by von Heijne and colleagues in this issue (Hessa *et al.*¹, page 377) tells us a lot more about what it takes to get into the membrane.

Membrane proteins are usually not soluble in the aqueous cytoplasm inside the cell. So cells have developed specialized machinery for injecting them into the membrane as they emerge from the complex in which they are synthesized. Hessa *et al.*¹ focused on the ubiquitous and best-studied insertion machine, known as the Sec translocon^{2,3}.

The Sec translocon must make several tricky decisions as the protein is being made; based on the emerging protein's amino-acid sequence, the translocon must choose

whether to pump the segment into the exoplasm outside the cell, push it sideways into the membrane, or flip it before releasing it into the membrane. Making the correct decision is essential if the protein is to fold properly, so the emerging peptide and the translocon must work in concert. Thus, to predict protein structure from amino-acid sequence, we need to understand how the translocon works with the nascent polypeptide to generate the final fold. The paper by Hessa *et al.*¹ helps to define how one of the decisions is made — how the translocon 'decides' whether to move a segment into the membrane or the exoplasm.

The cell membrane is a complex environment, and only specialized polypeptides, with equally complex and variable surface properties, reside there^{4,5}. For simplicity, we can divide the membrane into three regions (Fig. 1, overleaf): the central core of lipid hydrocarbon chains (about 30 Å thick); the interfacial region near the lipid head groups (about 15 Å thick); and the aqueous region. These environments are dramatically different, and so the structure of a membrane

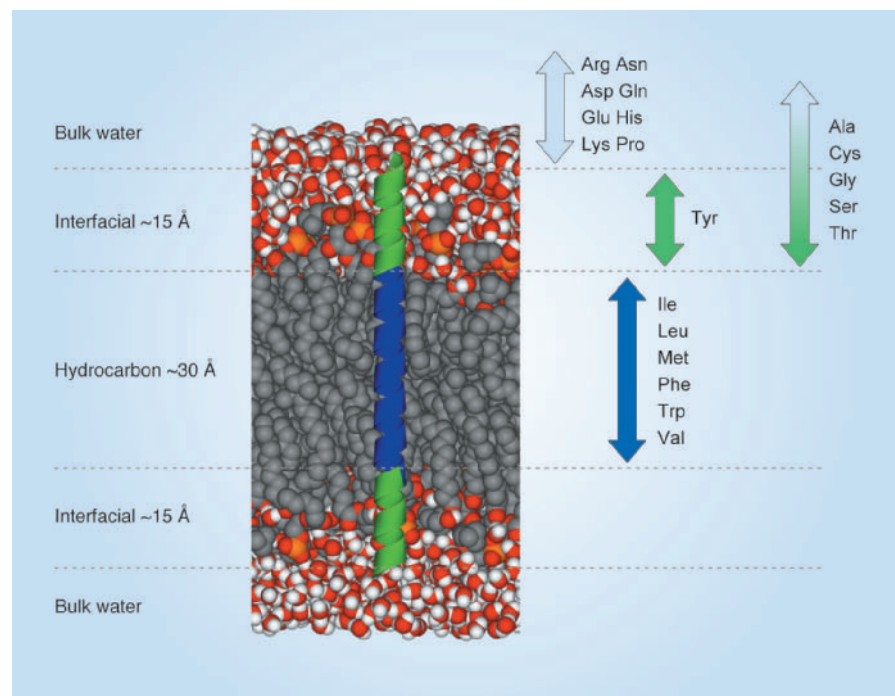


Figure 1 Membrane regions and preferred amino-acid locations. A snapshot of a lipid bilayer membrane¹¹ and its three major regions. Grey, carbon atoms; red, oxygen; white, hydrogen bound to oxygen; orange, phosphorus. In an α -helix, 20 amino acids (blue) can span the hydrocarbon core, and 10 amino acids (green) can span the interfacial region. Arrows indicate where most of each amino acid (denoted by its three-letter symbol) would be found at equilibrium based on transfer free-energy measurements^{7,8}.

protein needs to be such that each amino acid can interact favourably with its local environment. In previous studies^{6–8} to assess where in the membrane individual amino acids tend to go, Wimley and White measured the free energies required for transferring each of the 20 amino acids from water to either the interfacial region or the core region (summarized in Fig. 1). They found that amino acids containing apolar (hydrophobic) side chains prefer the hydrocarbon core, those that are moderately polar prefer the interfacial region, and the strongly polar amino acids prefer to be in water.

To assess the rules for membrane insertion by the Sec translocon, Hessa *et al.*¹ used

a rapid assay for distinguishing whether the translocon inserted a 19-amino-acid test segment into the membrane or pushed the segment across it. By varying the sequence of amino-acid residues in the test peptide, they could explore the determinants of the decision process. A 19-residue peptide in a helical conformation (the most common structure for traversing the membrane bilayer) is just large enough to span the membrane's hydrocarbon core (Fig. 1). According to the Wimley and White hydrophobicity scale^{6–8}, a 19-residue sequence consisting entirely of alanine amino acids (polyalanine) should prefer to remain outside the hydrocarbon region of the bilayer. To make it

sufficiently hydrophobic to favour insertion in the membrane, four or five leucine amino acids would need to be added.

There is no reason why the insertion machinery must be tied directly to the thermodynamics of membrane–water transfer, as it is not hard to imagine that an amino-acid segment could be trapped kinetically in the membrane. But, remarkably, the system behaves almost as though it were consulting the Wimley and White scale. When the test segment was entirely polyalanine, almost none was inserted into the membrane. But as more and more leucines were added, an increasing percentage of the test segment was inserted — when five leucines were present, more than 90% of the segment went into the membrane. This gives the impression that the translocon machinery somehow directs peptides according to a thermodynamic scale.

Strikingly, Hessa *et al.*¹ found that their data could be accurately modelled by imagining that an equilibrium between the translocated and inserted forms actually exists, providing a quantitative framework for their results. By replacing an amino acid within the test segment with each of the 20 possible amino acids, one at a time, they could measure apparent exoplasm–membrane transfer free energies for the insertion of each amino acid. This 'biological' scale correlates surprisingly well with the physical chemistry scale discussed above. The authors also examined how insertion probabilities vary according to the location of the amino acid in the test segment. As would be expected from the physical chemistry, amino-acid residues that favour the interfacial or aqueous regions of the membrane are more likely to be inserted in the membrane if they occur towards the end of the peptide. In addition, the peptide is more likely to be pushed through the membrane as it becomes more amphipathic (with one side of the helix being more polar than the other).

These results strongly suggest that the translocon decides whether to insert a segment into the membrane by measuring whether the segment best suits a hydrocarbon, interfacial or aqueous environment. How can it determine these free-energy values? One model (depicted in simplified form in Fig. 2) posits that the nascent polypeptide enters an aqueous translocon pore that can open towards the membrane core on one side^{2,9}. A potential lateral opening site is seen in the recent crystal structure of a related translocon¹⁰. Thus, the translocating polypeptide segment would find itself able to interact with both an aqueous and a bilayer environment. If the rate of translocation is slow enough relative to the rate of lateral opening and closing, it could be possible to establish an equilibrium between the aqueous and lipid environments. The more strongly the equilibrium favours the

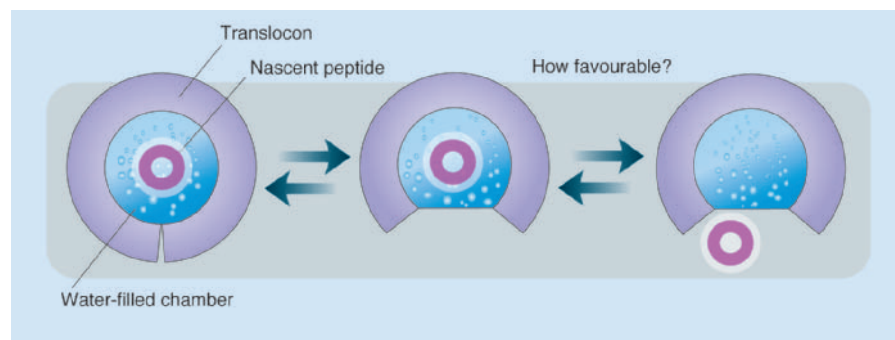


Figure 2 A possible scheme for the membrane-insertion decision, as proposed by Hessa *et al.*¹.

A top view of the membrane and the translocon pore that crosses it. Inside the pore is a peptide helix surrounded by water. The pore opens sideways into the membrane, allowing the helix to interact with the membrane lipids. If the peptide is more compatible with lipid than with water, it will transfer into the membrane; otherwise, it will continue to be moved through the pore. The figure is only intended to convey the basic principle, and omits many mechanistic and structural issues.

membrane environment, the higher the probability that the segment will stay there. Moreover, an amphipathic segment would find a favourable location in the lipid–water interface near the opening site, decreasing the likelihood that it will partition into the membrane.

Hessa and colleagues' results establish rules for membrane-protein insertion and folding. The thermodynamic scale derived in this work provides a starting point for understanding the preferences for the insertion of isolated helices. It seems likely that other factors will also contribute to this process, however. For example, interactions between neighbouring helices are likely to be a major factor. If a previously inserted helix interacts strongly with a translocating segment, it could drive the putative equilibrium in favour of membrane insertion, even if that segment would not normally insert in isolation. Loop lengths and loop folding could have a similar role. In addition, other proteins (such as TRAM²) could be involved in chaperoning the transmembrane segments into the membrane. Nevertheless, the work of Hessa *et al.* builds a quantitative

thermodynamic foundation that will allow these questions to be addressed. It also provides an important step towards the ultimate goal of predicting membrane-protein structure from sequence information. ■

James U. Bowie is in the Department of Chemistry and Biochemistry, Molecular Biology Institute, and the UCLA-DOE Laboratory of Genomics and Proteomics, University of California, Los Angeles, California 90095-1570, USA. e-mail: bowie@mbi.ucla.edu

1. Hessa, T. *et al.* *Nature* **433**, 377–381 (2005).
2. Rapoport, T. A., Goder, V., Heinrich, S. U. & Matlack, K. E. *Trends Cell Biol.* **14**, 568–575 (2004).
3. White, S. H. & von Heijne, G. *Curr. Opin. Struct. Biol.* **14**, 397–404 (2004).
4. White, S. H. & Wimley, W. C. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365 (1999).
5. White, S. H., Ladokhin, A. S., Jayasinghe, S. & Hristova, K. *J. Biol. Chem.* **276**, 32395–32398 (2001).
6. White, S. H. *FEBS Lett.* **555**, 116–121 (2003).
7. Wimley, W. C. & White, S. H. *Nature Struct. Biol.* **3**, 842–848 (1996).
8. Wimley, W. C., Creamer, T. P. & White, S. H. *Biochemistry* **35**, 5109–5124 (1996).
9. Heinrich, S. U., Mothes, W., Brunner, J. & Rapoport, T. A. *Cell* **102**, 233–244 (2000).
10. van den Berg, B. *et al.* *Nature* **427**, 36–44 (2004).
11. Tieleman, D. P., Sansom, M. S. & Berendsen, H. J. *Biophys. J.* **76**, 40–49 (1999).

Materials science

Build your own superlattice

Guus Rijnders and Dave H. A. Blank

Artificial materials made from oxide building blocks turn out to be excellent ferroelectrics. This shows that materials with specific properties can be designed by atomic-scale tailoring of their composition.

Ferroelectric oxides are used in a wide range of applications — random-access memories in computers, accelerometers in airbags or inkjet printers, telecommunication signal-processing devices and high-frequency devices for ultrasonic medical imaging, to name just a few. Predictions¹ that the performance of a ferroelectric oxide can be significantly improved by combining it with other oxides in a carefully tailored lattice have now been borne out by experiment. On page 395 of this issue, Lee *et al.*² show that such a 'superlattice' has a 50% enhancement in ferroelectric polarization compared with barium titanate, its only ferroelectric component. One of the key aspects of their method is the degree of control achieved at the atomic level during the growth of this artificial material.

Ferroelectrics are materials in which positive and negative charge centres separate spontaneously so that one side of the material is positive and the other negative. This polarization of charge exists even in the absence of an external electric field and is stable until an electric field is applied to change its direction. Device applications often make use of the piezoelectric properties

of ferroelectric materials; when a voltage is applied across a piezoelectric material, it undergoes a mechanical distortion in response and vice versa.

A potential application of ferroelectric materials lies in ultra-high-density memory devices, produced by controlling the ferroelectric domains at the nanometre scale³. Such storage devices would be non-volatile (and so able to retain the stored data for long periods of time without any power supply) and have short boot-up times.

Most materials used for ferroelectric devices are perovskites — oxides with a structure like that of the natural mineral CaTiO₃. Although the crystal structure of all perovskites is similar, their properties can differ significantly. For example, CaTiO₃ is a dielectric (resistant to electrical current), but replacing calcium with barium or lead produces piezoelectric materials. Partial substitution of titanium by zirconium in lead titanate gives lead zirconium titanate, at present the most widely used piezoelectric material.

Interest in perovskites received a boost with the discovery of high-temperature superconductivity in La–Ba–Cu-oxide in

1986 by Bednorz and Müller⁴. In the past decade, thin films of perovskites have been extensively studied to explore their electrical, magnetic and optical properties further. In the course of this work, progress has been made in controlling the growth of films at an atomic level, and Lee *et al.*² have built on several developments to make their superlattice structure — a stack of hundreds of thin perovskite layers.

To grow the films, Lee and co-workers used pulsed laser deposition, a technique that is particularly suitable for growing multi-component oxide structures. In this approach, a plasma is created by using laser pulses to evaporate oxide material from solid targets; this plasma has the same composition as that of the target. The approach can be used at relatively high oxygen pressures, which makes it possible to deposit stable units of perovskites under a wide range of conditions. To control the assembly of these units into a superlattice structure, Lee *et al.* make use of reflection high-energy electron diffraction.

When building up a superlattice, the layers must be grown carefully on top of each other so that their atomic lattices match. The termination (final atomic configuration) of each deposited layer will influence how well the layers grow, and consequently determine the device's performance. A prerequisite for controlling termination is an atomically flat substrate to start from. It was therefore a step forward when it became possible to prepare substrates that are terminated in a single configuration, and so are atomically flat^{5,6}. Lee and colleagues need a conducting electrode, and use SrRuO₃, a metallic ferromagnetic perovskite, as a substrate. It can be grown atomically smooth, with a termination of SrO (refs 7, 8).

Previous work showed that superlattices can be designed with specific properties; for example, neither BaCuO₂ nor SrCuO₂ exhibits superconducting behaviour, whereas a superlattice consisting of thin layers of both oxides does so⁹. Another example is the superlattice of SrZrO₃/SrTiO₃. Neither of the two building blocks is ferroelectric, but the superlattice is¹⁰.

Lee *et al.* assemble their superlattice with three different building blocks — BaTiO₃, SrTiO₃ and CaTiO₃ (Fig. 1, overleaf). The use of three different compounds breaks the inversion symmetry that often occurs in two-component superlattices¹¹. Typically, ferroelectric materials display symmetric two-state polarization (that is, the applied electric field required to change it in either direction is equal). In the three-component superlattice, inversion symmetry is broken, resulting in asymmetric polarization and an extra degree of freedom for optimizing the ferroelectric properties.

Growing thin layers on top of each other can lead to considerable 'epitaxial' strain in

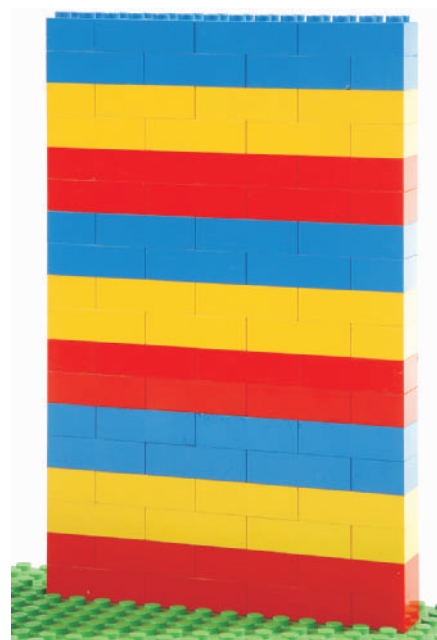


Figure 1 Building regulation — a Lego version of the superlattice structure shown in Fig. 1d (page 396)². The base constitutes the substrate, including the SrRuO_3 electrode, and the Lego wall is made of layers of three different bricks: SrTiO_3 (red), BaTiO_3 (yellow) and CaTiO_3 (blue). Lee *et al.*² demonstrate that an artificial material with favourable ferroelectric properties can be constructed by using these three perovskite building blocks.

the structure, which is caused by differences between the lattice parameters of the layers. This effect influences the properties of the layers and the structure as a whole, and Lee *et al.* have exploited this phenomenon to increase the ferroelectric polarization of their material. The epitaxial strain decreases the in-plane lattice parameters of the BaTiO_3 layers, producing an increase in the ferroelectric polarization.

Lee and colleagues' ferroelectric superlattice is impressive, because it shows that such structures can be built with atomic precision and possess properties that surpass those of the individual building blocks. It also underlines the potential of designing artificial superlattices with unique properties. For example, most ferromagnetic materials show no ferroelectricity. But what would a superlattice made from ferroelectric and ferromagnetic building blocks be like? Would it show ferroelectric or magnetic behaviour, or have a mix of properties? New phenomena might even emerge. Here, then, is an invitation to materials scientists to design and build their own superlattice. ■

Guus Rijnders and Dave H. A. Blank are at the MESA+ Institute for Nanotechnology, and the Faculty of Science and Technology, University of Twente, PO Box 217, Enschede, The Netherlands.

e-mail: d.h.a.blank@utwente.nl

1. Neaton, J. B. & Rabe, K. M. *Appl. Phys. Lett.* **82**, 1586–1588 (2003).
2. Lee, H. N., Christen, H. M., Chisholm, M. F., Rouleau, C. M. & Lowndes, D. H. *Nature* **433**, 395–399 (2005).
3. Ahn, C. H., Rabe, K. M. & Triscone, J. M. *Science* **303**, 488–491 (2004).
4. Bednorz, J. G. & Müller, K. A. Z. *Phys. B* **64**, 189–197 (1986).
5. Kawazaki, M. *et al. Science* **226**, 1540–1542 (1994).
6. Koster, G. *et al. Appl. Phys. Lett.* **73**, 2920–2921 (1998).
7. Rijnders, G., Blank, D. H. A., Choi, J. & Eom, C. B. *Appl. Phys. Lett.* **84**, 505–507 (2004).
8. Lee, H. N. *et al. Appl. Phys. Lett.* **84**, 4107–4109 (2004).
9. Norton, D. P. *et al. Science* **265**, 2074–2077 (1994).
10. Tsurumi, T. *et al. Appl. Phys. Lett.* **85**, 5016–5018 (2004).
11. Warusawithama, M. P. *et al. Phys. Rev. Lett.* **90**, 36802 (2003).

Ecology

Paradise sustained

Shahid Naeem and Andrew C. Baker

Biodiversity stabilizes ecosystem functioning in small-scale, short-term experiments, but do such findings scale up to the larger world? A global study of fossil reefs from the past 500 million years suggests they do.

A watershed ecosystem that produces a steady volume of water may be more valuable than one that unpredictably alternates between flood and drought. A coastal ecosystem that provides a regular but modest supply of fish serves a community better than one that booms then busts. In most cases, then, the absolute magnitudes of ecosystem functions — such as the production of potable water and nutritious food — may matter little if they are unsustainable.

Over the past decade, some of the most intense research ever conducted in ecology has examined how the magnitudes and sustainability of ecosystem functions are governed by biodiversity, and whether a loss of biodiversity is a major cause of ecosystems becoming less productive and less stable^{1–4}. In this debate, the 'ayes', who advocate a strong role for biodiversity in governing ecosystem function, point to the congruency of findings from the full triad of scientific methods (theory, observation and experiment). The nay-sayers find the evidence flimsy — weak effects derived from studies too small and too short-lived to be convincing.

On page 410 of this issue, Kiessling⁵ provides evidence that may sway some nay-sayers — but there's a catch. Making use of an extensive palaeoecological database on thousands of reefs, spanning some 500 million years, Kiessling tested whether the local species richness of reef-builders on a given reef (and, by extension, the diversity of the entire community) could predict ecological change on these reefs. Using reef type, density, architecture and construction style as proxies for reef ecology, Kiessling found that higher-than-average species diversity in one time interval led to lower-than-average changes in reef ecology in the next. In other words, biodiversity may indeed govern sustainability. The catch is that this epoch-spanning data set cannot detect fluctuations in reef ecology of less than a few million years, which may disappoint some ecologists and ecosystems biologists following the debate.

For the ayes, the good news is that the positive biodiversity–stability relationship

predicted from theoretical and empirical work may scale up — way up. Although some may quibble with Kiessling's use of both absolute and relative measures of species richness in his analyses, and others may question whether the variables he chose are appropriate measures of reef ecology, no one is likely to dispute the depth and scope of his perspective. Kiessling's study is global and covers an extraordinary length of time, one that spans three entire geological eras and witnesses the wholesale turnover of reef-building consortia, from cyanobacteria to calcified sponges, bryozoans, molluscs, calcareous algae, and corals both ancient and modern (Fig. 1). Given this, it is remarkable that Kiessling's findings agree with those derived from mathematical models, bottles, Petri dishes, aquaria, growth chambers and field plots, which often run for only a handful of generations.

However, the 'deep time' approach to the diversity–stability debate is not without pitfalls. In Kiessling's study, some data points that documented major ecological change following periods of relatively high biodiversity were excluded from the analysis. These data, which contradict the diversity–stability hypothesis, are from three out of the five major mass-extinction events covered by the study.

Kiessling rightly argues that some external influences (such as asteroid impacts) really are beyond the scope of biology to buffer. However, in recognizing and excluding these legitimate outliers, we should take care not to throw the evolutionary baby out with the ecological bathwater. Many of these mass-extinction outliers are superimposed on a backdrop of major evolutionary milestones, such as the rise of predation and herbivory as lifestyle strategies, and the acquisition of algal photosymbionts in today's stony corals⁶ — events that are themselves likely to result in dramatic ecological shifts in reef communities. Resolving this debate over extended timescales will therefore require us to distinguish between ecology and evolution in determining stability and

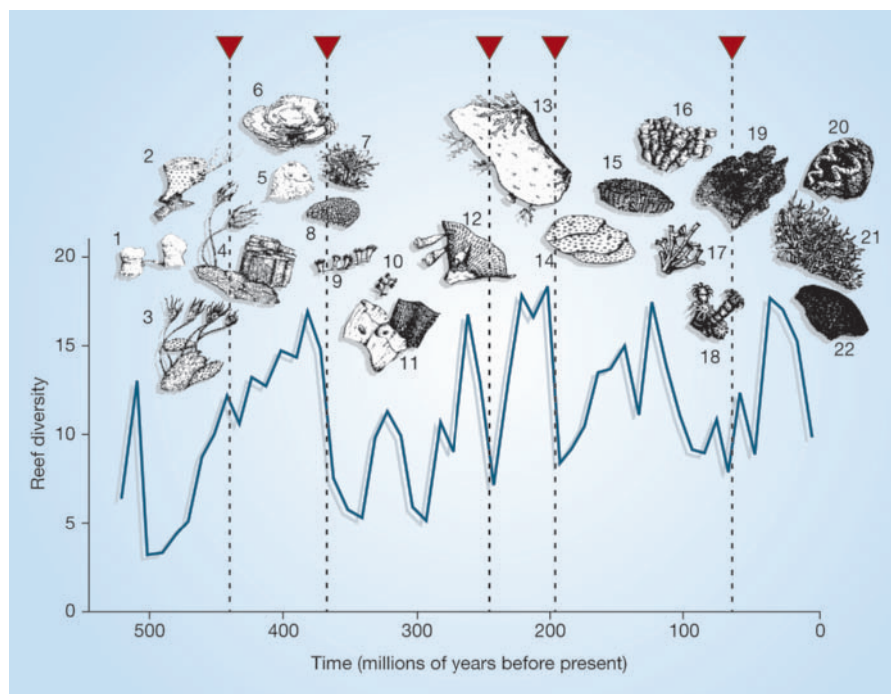


Figure 1 Reef-builders: diversity and stability. Over the 500-million-year time frame studied by Kiessling⁵, the diversity of species in coral reefs has fluctuated, and the dominant reef-building organisms have included cyanobacteria, sponges, molluscs, algae and corals. Kiessling used many ecological indicators to show that periods of high biodiversity are followed by periods of lower-than-average ecological change — in other words, biodiversity leads to stability. The red arrowheads represent times when major extinctions occurred. Diagrams drawn by John Sibbick, from ref. 6. Graph from ref. 5, supplementary information. Stromatolites: 1, 5; sponges: 2, 5–7, 12, 13; crinoids: 2–4; cyanobacteria: 2, 3, 6, 7; tabulate corals: 3, 4, 8, 10; rugose corals: 9; foraminifera: 11; bryozoans: 11, 12; rudist bivalves: 16; algae: 11, 12; scleractinian corals: 14, 15, 17, 18, 21, 22; gorgonian coral: 19; giant clam: 20. Organisms not drawn to scale.

productivity. Either evolutionary innovation drives ecological change, and the coupling of biodiversity and stability are ancillary, or the two are so tightly linked that it takes mass extinctions and dramatic evolutionary innovation to decouple them.

These ecological and evolutionary findings⁵ are extraordinarily important to the diversity–stability discussion. Those who follow the debate because of its contemporary environmental implications, however, may find this study's multi-million-year time frame frustrating—but there is an important message. The latest rounds in the biodiversity–stability debate have been fuelled by studies of current rates of species loss and their impacts on current ecosystem function. Kiessling's finding that, given a few million years, species-rich ecosystems will tend to look the same as they do now, won't do much to settle this argument. Indeed, given that today's reefs are relatively diverse, one might inadvertently conclude that they will persist in the face of environmental perturbation.

Kiessling's data, however, show that only 5–58% of ecological change is predicted by diversity, leaving plenty of scope for environmental impacts to have a dramatic effect. If history provides some basis for predicting the future, coral reefs may show less variability over the next few million years than their

less-diverse predecessors, but they might also collapse relatively easily. Threatened by climate change and coral bleaching, as well as by overfishing, the spread of marine diseases, nutrient pollution and habitat destruction⁷, reefs lie at the threshold of these alternative futures. The diversity of coral reefs may make them productive, and it may sustain their function over time, but these reef paradises can still be lost in the face of environmental change.

Shahid Naeem and Andrew C. Baker are in the Department of Ecology, Evolution and Environmental Biology, Columbia University, 1200 Amsterdam Avenue, New York, New York 10027, USA. Andrew C. Baker is also in the Marine Program, Wildlife Conservation Society, 2300 Southern Boulevard, Bronx, New York 10460, USA.
e-mail: sn2121@columbia.edu

1. McCann, K. S. *Nature* **405**, 228–233 (2000).
2. Hughes, J. B., Ives, A. R. & Norberg, J. in *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives* (eds Loreau, M., Naeem, S. & Inchausti, P.) 92–101 (Oxford Univ. Press, 2002).
3. Loreau, M. et al. in *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives* (eds Loreau, M., Naeem, S. & Inchausti, P.) 79–91 (Oxford Univ. Press, 2002).
4. Norberg, J. et al. *Proc. Natl Acad. Sci. USA* **98**, 11376–11381 (2001).
5. Kiessling, W. *Nature* **433**, 410–413 (2005).
6. Wood, R. *Reef Evolution* (Oxford Univ. Press, 1999).
7. Hughes, T. P. et al. *Science* **301**, 929–933 (2003).



100 YEARS AGO

"Compulsory Greek at Cambridge." When I decided to go up to Cambridge to study mathematics and philosophy I was living abroad, and I crammed Greek... But on going in for the "Little Go," though I passed easily in translation, I failed by a few marks in Greek grammar. It was so near a thing that I thought I might pull through in December with a few hours more grind; but unfortunately I ran it too fine, and again failed by a few marks. This meant that I had to get up a complete new set of translation books for the following June, and to prevent further mistakes I went to a coach for the grammar part. I then passed, getting a second class... I can only say my present knowledge of the language is *nil*, although I had a double dose of it... But this may have been due to my resentment at being forced to waste time in an uncongenial study, when I was keen to get on to something else.

Edward T. Dixon

From *Nature* 26 January 1905.

50 YEARS AGO

The declared aim of the Central African Federation is a partnership between the races, European and non-European; and to the attainment of this ideal a university, with the goodwill of the people and the support of the government, can make a tremendous contribution... The Moffat Resolutions adopted by the Legislative Council of Northern Rhodesia on July 29, 1954 (*The Times*, August 14, 1954), began the task of defining partnership in a more precise way than has hitherto been attempted: "(1) The objective of policy in Northern Rhodesia must be to remove from each race the fear that the other might dominate for its own racial benefit and to move forward from the present system of racial representation in the territorial legislature towards a franchise with no separate representation for the races. (2) Until that objective can be fully achieved a period of transition will remain during which special arrangements in the Legislative and Executive Councils must continue to be made so as to ensure that no race can use either the preponderance of its numbers or its more advanced stage of development to dominate the other for its own racial benefit..." The significance of education, and particularly the education of the African, in the working out of such a plan must be clear to all.

From *Nature* 29 January 1955.

Obituary

Maclyn McCarty (1911–2005)

Maclyn McCarty, physician and microbiologist, died of congestive heart failure in New York City on 2 January. He was the last survivor of the three-man team that demonstrated that genes are made of DNA.

Previously, most scientists believed that genes must be made of proteins and that DNA lacked the necessary complexity to carry hereditary information. But in 1944, Oswald T. Avery, Colin MacLeod and McCarty published the first paper, in a series of three, that conclusively showed DNA to be the carrier of genetic information. This paper was entitled “Studies on the chemical nature of the substance inducing transformation of pneumococcal types”, and the research crowned a 30-year effort, undertaken in Avery’s laboratory at Rockefeller Hospital in New York, to understand and combat pneumococcal pneumonia.

The trio’s discovery had its origins in the research of an Englishman, Fred Griffith, who in 1928 described the phenomenon of the transformation of pneumococcal types. All of Griffith’s work was done *in vivo*, by manipulating pneumococcal bacteria in mice. Transformation involved a heritable change in one type of pneumococcus that was induced by a substance extracted from another type. The change was predictable and permanent, and could be transmitted from generation to generation in laboratory-grown organisms. This finding had, in McCarty’s phrase, “all the earmarks of what we would call today, the transfer of genetic information”.

We now know that new microbes and new infectious diseases may emerge as a result of genetic mechanisms in pathogens and selective evolutionary pressures. Yet, strange as it may seem, microbial genetic evolution is a new idea that became established long after scientists understood genetics in plants and animals. So it is no surprise that the 1944 paper was greeted with some scepticism. Nevertheless, the paper had an immediate and profound influence on the thinking and the research direction of many scientists. Frank Macfarlane Burnet, for instance, later wrote, “the discovery that DNA could transfer genetic information from one pneumococcus to another heralded the opening of the field of molecular biology”. McCarty recounted the history of the discovery in his memoir *The Transforming Principle*.

According to McCarty’s mother,



Co-discoverer of DNA as the material of heredity

medicine and medical research were his ambition by the time he was ten years old. While in high school, his goal was to enter Johns Hopkins School of Medicine. First at Stanford University, then at Johns Hopkins, McCarty excelled at biochemistry. After medical school he was a paediatric house officer for three years, having selected paediatrics “as the best way to promote my interest in infectious diseases”.

Later, following his research on pneumococcal transformation and heredity, McCarty remained at the Rockefeller University and was appointed director of the Laboratory of Streptococcal Infections and Rheumatic Fever — the latter being a devastating childhood disease at the time. Like Avery, McCarty accepted only several junior colleagues at a time. But we were a privileged few. In an age of big science, McCarty’s laboratory was a cottage industry.

McCarty used the newly emerging techniques and concepts of cell biology in his approach to streptococcal research, delineating the chemical architecture of these bacterial cells and identifying the major streptococcal antigens as constituents of the cell wall. In great detail, McCarty studied the chemistry of the cell-wall ‘A carbohydrate’. He surmised that information about the structure of this substance might yield an explanation for the pathogenesis of rheumatic fever. McCarty never put into print his speculation that the immunological crossreactivity between the A carbohydrate and antigens in human connective tissue was related in some way to this pathogenesis. As Avery often said,

“stick to the facts” — and McCarty did. The crossreactivity hypothesis is still not proven, but as McCarty recently noted, “it seemed reasonable to pursue at the time”.

From 1954, McCarty and his colleagues began intensive studies on the viruses (bacteriophages) that infect streptococci. All efforts to transform group A streptococci had failed. But there was reason to believe that research on bacteriophages would allow entry into the genetics of streptococci and, in particular, into the genetic basis of the toxic substances they produce and the clinical conditions they cause. That expectation has since been proven correct by numerous investigators. We now know, for example, that bacteriophages carry the genes that encode the toxins responsible for scarlet fever. McCarty’s work also helped pave the way for genetics research on many other pathogens, including *Vibrio cholerae* and staphylococci, and to further studies of phage-encoded toxins and other microbiological phenomena.

Although McCarty retired in 1981, he continued to go into the laboratory until the end, to discuss research progress with junior colleagues. Retirement also gave him and his wife Marjorie time to travel widely, particularly in France, England and Germany, where they had many friends who treasured their companionship. Often, they rented an apartment in Paris and settled into a daily routine. McCarty read French fluently, but he was reticent about speaking the language. In this regard Marjorie was not so shy, and she would also scan the *Michelin* guide for events that they would attend and for starred restaurants where they would dine. Even in France, dinner was preceded by a Tanqueray Gibson on the rocks. In earlier years, all of us had a Gibson or a gin Martini straight-up. MacLeod related that Avery would announce at the end of a long day in the lab: “Now let’s go have ‘something else’” — his term for a dry gin Martini.

The 1944 paper on pneumococcal transformation begins: “Biologists have long attempted by chemical means to induce in higher organisms predictable and specific changes which thereafter could be transmitted in series as hereditary characters.” All that, and then some, has come to pass: witness the human genome, recombinant DNA technology, and genetically engineered animals that produce complex proteins such as human antibodies.

Richard M. Krause

Richard M. Krause is at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 16 Center Drive, Bethesda, Maryland 20892, USA.
e-mail: richard_krause@nih.gov

Cancer

Telomeres on a short leash

Cancer Cell **7**, 25–37 (2005)

Treatments that throw a double punch at chromosome ends could kill off cancer cells, Hiroyuki Seimiya *et al.* suggest. DNA sequences known as telomeres, typically found at chromosome ends, help normal cells keep tabs on how many times they have divided. Chromosomes must be replicated before cells can divide, but the very tips of telomeres cannot be copied, so, with each division, these sequences get shorter and shorter. Eventually, they become so truncated that the DNA cannot be properly replicated, and the cell stops dividing.

Cancer cells, however, contain an activated enzyme called telomerase, which prevents telomeres from shortening and leads to uncontrolled proliferation. So researchers have been developing treatments that inhibit this enzyme. But these therapies do not always succeed, and can produce drug resistance.

To develop a better treatment option, Seimiya *et al.* investigated what happened when they inhibited both telomerase and another enzyme, tankyrase 1, that helps it gain access to telomeres. This enhanced the shortening of these chromosomal ends, and accelerated cancer-cell death. The authors suggest that tankyrase 1 inhibitors might significantly boost the efficacy of telomere-targeted therapies.

Roxanne Khamsi

Developmental biology

Switch of fate

Cell **120**, 123–135; 137–149 (2005)

Neurons produce two very distinct types of projection: a single axon, which passes signals on to neighbouring cells, and numerous dendrites, which receive signals from axons. Two groups have now found that an enzyme called glycogen synthase kinase-3 β (GSK-3 β) helps to determine what a projection becomes.

Hui Jiang *et al.* and Takeshi Yoshimura *et al.* showed that GSK-3 β blocks the formation of axons. When they boosted the activity of GSK-3 β in cultured neurons, the cells often lacked axons completely. Conversely, when GSK-3 β was inhibited, the cells sprouted several axons — or converted existing dendrites into axons.

The groups went on to identify molecules working upstream and downstream of GSK-3 β . They showed that GSK-3 β is tagged with phosphate groups and thereby activated by the enzyme Akt. GSK-3 β then itself phosphorylates and inhibits a molecule called CRMP-2, which normally promotes the assembly of the neuron's internal 'skeleton' and hence axon growth.

Self-assembly

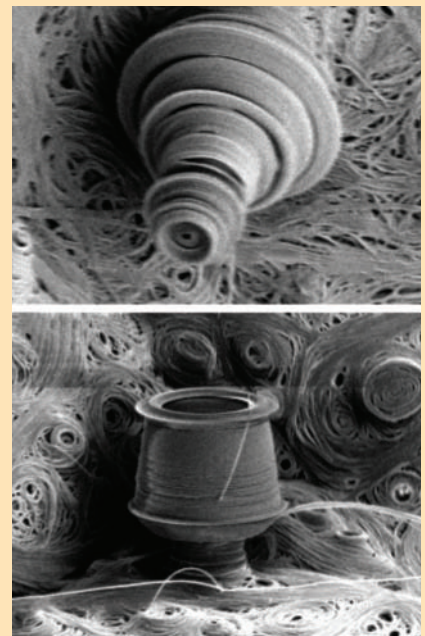
Pots of soap

Langmuir **21**, 516–519 (2005)

Structures that look like earthenware spun on a wheel emerge spontaneously from a solution of a common cosmetics surfactant, Janhavi S. Raut *et al.* report. This 'micro-pottery' is built from aggregates of the surfactant sodium myristate, which self-assemble when a concentrated solution is cooled to room temperature.

For fast cooling rates, the fibres simply tangle into a gel-like mat. If cooling is slower than 2 °C per minute, the fibres begin to curl into rings; at 1 °C per minute they form bundles like thread wound on a bobbin. Air bubbles in the cooling solution become covered in bundles in which the fibres are spooled into appealing shapes (pictured), about 10–20 μ m across.

Raut *et al.* think that the initial bending of the fibres into rings, which act as seeds for 'spinning' the micro-pots, is triggered by pressure gradients at the air–water interface of the bubbles. These gradients arise from differences in surfactant concentration across the surface, leading to variations in surface tension that drive liquid flow: the so-called Marangoni effect. The



micro-pots represent a third level of structural hierarchy in the surfactant assemblies, which form sheets that then gather in layers to create the ribbon-like fibres.

Philip Ball

The discovery raises the possibility that drugs that block GSK-3 β could be used to promote axonal growth and so enhance the repair of neuronal injuries, although the researchers have not yet shown that the new axons would work normally in animals.

Helen Pearson

planets themselves. Further study will focus on the size and composition of the particles in the disk.

Mark Peplow

Zoology

Estimating the unknown

Proc. R. Soc. Lond. B doi:10.1098/rspb.2004.2955

Andrew R. Solow and Woollcott K. Smith have taken a fresh look at estimating the number of species in a particular category of animals, including those that are yet to be discovered.

A traditional approach is to take the cumulative historical record of discoveries of species, and extrapolate into the future. Solow and Smith propose a statistical model for the 'discovery process', which can be fitted to the data using the principle of maximum likelihood, based in probability theory. Central elements factored into the procedure are the animals' visibility, and a function to account for changes over time in the skill and effort applied in sighting new species.

The authors test their model against a previously published record of the dates when large marine animals (defined as 2 metres or more in length) were discovered, starting in 1828 and ending in 1996. That record comprises 117 species, with a further 100 having been identified before 1828.

The upshot of the calculations, say Solow and Smith, is that around 10 large marine animals remain to be discovered, and probably 16 at most. The most likely candidates are fish and cephalopods.

Tim Lincoln

Astronomy

Dusty dwarfs

Astrophys. J. (in the press)

Brown dwarfs are the failed stars of the cosmos. With insufficient mass to fuse hydrogen atoms, they glow weakly by burning deuterium. Astronomers are still unsure whether the smallest brown dwarfs form like their larger cousins — by accreting matter from a swirling disk of dust particles.

K. L. Luhman and colleagues have found the smallest and coolest brown dwarf to date that is surrounded by a dusty disk. The dwarf, called OTS 44, is about 15 times the mass of Jupiter, right at the lower mass limit necessary for deuterium burning. The observation of a disk around such a small star implies that the accretion mechanism works for even the smallest stars.

The researchers used infrared measurements from the Spitzer Space Telescope to identify the disk around the dwarf star. It lies in the Chamaeleon I cloud complex, a nearby star-forming region about 554 light years from Earth.

The finding raises the possibility that planets could form from dust around brown dwarfs that have barely more mass than

Rapid speciation in an arthropod

The likely force behind an explosion of new Hawaiian cricket species is revealed.

Theory predicts that sexual behaviour in animals can evolve rapidly, accelerating the rate of species formation^{1,2}. Here we estimate the rate of speciation in *Laupala*, a group of forest-dwelling Hawaiian crickets that is characterized primarily through differences in male courtship song³. We find that *Laupala* has the highest rate of speciation so far recorded in arthropods, supporting the idea that divergence in courtship or sexual behaviour drives rapid speciation in animals.

Identifying groups of organisms characterized by high rates of speciation can reveal factors that promote speciation. We used amplified fragment-length polymorphisms (AFLPs) to estimate a phylogenetic tree of *Laupala* (Fig. 1; for methods, see supplementary information). The topology of the tree and precise dating of the Hawaiian islands⁴ enabled us to estimate rates of speciation.

The highest speciation rate, 4.17 species per million years, was found in a monophyletic clade from Hawaii Island (Fig. 1). This rate is more than an order of magnitude greater than the average estimated rate of arthropod speciation, calculated as 0.16 per million years⁵, and is exceeded only by that of the rapidly speciating African cichlid fish⁶.

Our results also reveal a successive reduction in speciation rates with clade age (to 1.26 Myr⁻¹ and then 0.80 Myr⁻¹; see supplementary information). Such a decline is expected as older lineages suffer a greater amount of extinction, reducing the apparent number of species formed. Moreover, diversity in island lineages may be limited by island size⁷, with speciation rates on older islands declining as islands reach maximum species capacity. In the Hawaiian islands, this effect is magnified as islands have eroded and subsided, becoming smaller over time⁴.

By contrast, speciation on the actively growing and youngest island, Hawaii, seems to be in progress, as diversity in *L. cerasina* (Fig. 1) was probably masked by sampling a single population per species. Populations of *L. cerasina* are acoustically diverse³, and more intensive analysis using AFLPs reveals several distinct genetic groups that correspond to acoustic and geographical variations^{8,9}. We conclude that speciation on Hawaii Island is both explosive and ongoing.

Although the causes of speciation are difficult to ascertain experimentally, traits that distinguish closely related species provide important insight. In *Laupala*, closely related species are morphologically cryptic and distinguishable only in the pulse rate of their male courtship song, a secondary sexual trait used in mate attraction. Females prefer pulse rates of their own species¹⁰, so divergence in

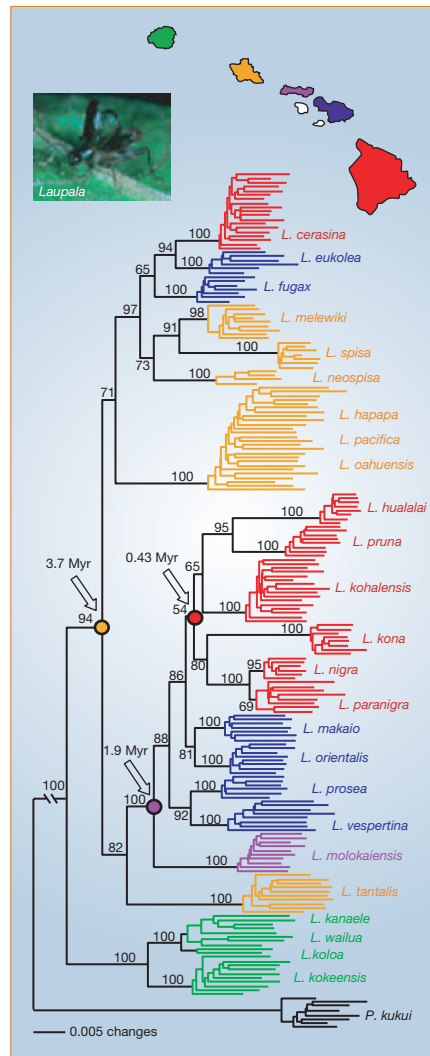


Figure 1 Phylogeny estimate based on analysis of amplified fragment-length polymorphisms for 25 of 38 described species of *Laupala* cricket. Terminal taxa are individuals; branches are colour-coded to indicate Hawaiian island of origin (green, Kauai; yellow, Oahu; purple, Molokai; dark blue, Maui; red, Hawaii). *Prolaupala kukui*, a member of the hypothesized sister genus to *Laupala*³, was used as the outgroup. Arrows indicate the ages of three most recent common ancestors used to estimate rates of speciation. For further details, see supplementary information.

male song reduces the chances of interbreeding between species.

Closely related species of *Laupala* have no ecologically distinguishable features³: they are dietary generalists, without host-plant dependency, and thrive in both native and introduced Hawaiian forests^{3,11}. They are frequently syntopic and synchronic, with males of different species often singing next to one another in the wild. It is therefore likely that the forces responsible for speciation in *Laupala* are those that cause the evolutionary divergence of secondary sexual traits.

If secondary sexual traits can evolve rapidly owing to sexual selection, then they could lead to accelerated rates of speciation^{1,2}. African cichlid fish exemplify this process, with sister species differing primarily in male coloration, a secondary sexual trait. From this pattern, it has been argued that the spectacular diversification in African cichlid fish is driven by sexual selection¹².

Whether sexual selection has promoted the diversification in song in *Laupala* remains to be proved, and reinforcing selection on the mating system is a possibility; however, other mechanistic explanations can reasonably be ruled out. Ecological speciation, for example, occurs when isolated populations evolve traits that allow them to exploit resources in new environments¹³. By this mechanism, natural selection drives speciation, producing species that differ in ecological traits adapted to novel environments. In *Laupala*, such ecological characters do not distinguish species, so this model is unlikely to explain speciation in this group.

We have shown that the highest rate of speciation recorded so far in arthropods belongs to a group of crickets that differ primarily in secondary sexual traits, indicating that divergence in sexual behaviour may cause this rapid speciation^{1,2}. AFLP technology allows high resolution of species relationships and unprecedented insight into the early stages of speciation. The best context in which to investigate the origin of species is provided by the early stages of speciation, before subsequent diversification.

Tamra C. Mendelson*†, Kerry L. Shaw*

*Department of Biology, University of Maryland, College Park, Maryland 20742, USA

†Present address: Department of Biological Sciences, Lehigh University, Bethlehem, Pennsylvania 18015, USA

e-mail: tamram@lehigh.edu

- West Eberhard, M. J. *Q. Rev. Biol.* **58**, 155–183 (1983).
- Pomiankowski, A. & Iwasa, Y. *Proc. Natl Acad. Sci. USA* **95**, 5106–5111 (1998).
- Otte, D. in *The Crickets of Hawaii: Origin, Systematics, and Evolution* (Orthoptera Society/Academy of Natural Sciences of Philadelphia, 1994).
- Carson, H. L. & Clague, D. A. in *Hawaiian Biogeography: Evolution on a Hot Spot Archipelago* (eds Wagner, W. L. & Funk, V. A.) 14–29 (Smithsonian Institution Press, Washington DC, 1995).
- Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).
- McCune, A. R. in *Molecular Evolution and Adaptive Radiation* (eds Givnish, T. J. & Sytsma, K. J.) 585–610 (Cambridge Univ. Press, New York, 1997).
- Losos, J. & Schluter, D. *Nature* **408**, 847–850 (2000).
- Parsons, Y. M. & Shaw, K. L. *Mol. Ecol.* **10**, 1765–1772 (2001).
- Mendelson, T. C., Siegel, A. M. & Shaw, K. L. *Mol. Ecol.* **13**, 3787–3796 (2004).
- Mendelson, T. C. & Shaw, K. L. *Genetica* **116**, 301–310 (2002).
- Shaw, K. L. in *Hawaiian Biogeography: Evolution on a Hot Spot Archipelago* (eds Wagner, W. L. & Funk, V. A.) 39–56 (Smithsonian Institution Press, Washington DC, 1995).

12. Knight, M. E. & Turner, G. F. *Proc. R. Soc. Lond. B* **271**, 675–680 (2004).

13. Schluter, D. *Trends Ecol. Evol.* **16**, 372–380 (2001).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Oscillatory motion

Quantum whistling in superfluid helium-4

Fundamental considerations predict that macroscopic quantum systems such as superfluids and the electrons in superconductors will undergo oscillatory motion when forced through a small constriction. Here we induce these oscillations in superfluid helium-4 (^4He) by pushing it through an array of nanometre-sized apertures. The oscillations, which are detected as an audible whistling sound, obey the so-called Josephson frequency relation and occur coherently among all the apertures. The discovery of this property in ^4He at the relatively high temperature of 2 K (2,000 times higher than the temperature at which a related but different phenomenon occurs in ^3He) may pave the way for a new class of practical rotation sensors of unprecedented precision.

The Josephson effects in superconductors have received attention both as an aid to scientific understanding and for their technological importance¹. Analogous effects, including Josephson oscillations, have been observed^{2,3} in superfluid ^3He below 1 mK. However, detection of oscillations at the Josephson frequency in superfluid ^4He has remained elusive until now, despite almost four decades of attempts⁴.

Superconductors and superfluids are both described by a macroscopic wave function that includes amplitude and phase, ϕ . A chemical-potential difference, $\Delta\mu = \mu_2 - \mu_1$, between two baths of superfluid separated by an aperture causes the phase difference, $\Delta\phi = \phi_2 - \phi_1$, to change in accordance with the Josephson–Anderson phase-evolution equation

$$\frac{d\Delta\phi}{dt} = \frac{-\Delta\mu}{\hbar}$$

where $\hbar$ is Planck's constant (h) divided by 2π and where $\Delta\mu/m_4 = \Delta P/\rho - S\Delta T$ (and m_4 is the mass of the ^4He atom, ΔP is the pressure difference, ρ is the mass density, S is the entropy per unit mass, and ΔT is the temperature difference). A non-zero $\Delta\phi$ results in a superfluid current, $I(\Delta\phi)$, through the aperture. If $I(\Delta\phi)$ is periodic for 2π , a constant $\Delta\mu$ causes current to oscillate through the aperture at the Josephson frequency $f_j = \Delta\mu/h$. The periodicity in $I(\Delta\phi)$ can occur if the aperture acts like an ideal weak link^{3,5}, in which case $I(\Delta\phi) \propto \sin(\Delta\phi)$, or by the generation of 2π phase slips⁶, in which case $I(\Delta\phi)$

is expected to follow a sawtooth waveform.

The experimental set-up is shown in Fig. 1a (for methods, see supplementary information). We used an electrostatically driven diaphragm² to apply an initial pressure step between two baths of superfluid separated by an aperture array. The array consisted of 65×65 nominally 70-nm apertures spaced on a $3\text{-}\mu\text{m}$ square lattice in a 50-nm-thick silicon nitride membrane. After the pressure step, fluid flowed through the array and the chemical-potential difference relaxed to zero. When the output of a diaphragm position sensor, which monitored fluid flow, was connected to a set of headphones, we heard a clear whistling sound that passed from high to low frequency (audio recording in supplementary information).

By using Fourier transform methods, we extracted the frequency and amplitude of this whistle as a function of time throughout the transient. Immediately after the pressure step is applied, the temperatures on either side of the aperture array are equal and the entire $\Delta\mu$ is determined by the initial pressure head, ΔP_0 . Figure 1b shows that the initial frequency is proportional to the initial chemical-potential difference. The slope of the line agrees, within the systematic error of our pressure calibration, with the Josephson frequency formula ($f_j = m_4\Delta P_0/\rho h$).

Oscillations resulting from 2π phase slips are expected to have a velocity amplitude $\kappa/2l$, where $\kappa = h/m_4$ is the circulation quantum and l is an effective length for one aperture⁷. If, in addition, the oscillation in each of the N apertures occurs coherently, the amplitude of the diaphragm-displacement Fourier component at f_j is

$$X_0 = \alpha \frac{\rho_s N \kappa a}{4\pi f_j \rho A l}$$

where A is the area of the diaphragm, a is the area of a single aperture, and ρ_s is the superfluid density. The factor α would be $2/\pi$ for a sawtooth waveform, or unity for a sinusoid of the same peak amplitude. We find $\alpha \approx 0.6$, independent of temperature in the range where, if T_λ is the superfluid transition temperature, $T_\lambda - T$ is between 1.7 and 2.9 mK.

We conclude that the oscillation is a coherent phenomenon involving all the apertures in the array, and is possibly sawtooth in waveform. This coherence is remarkable, because earlier work using a single aperture showed that thermal fluctuations in the phase-slip nucleation process destroy time coherence in the rate of phase slippage, so that no Josephson oscillation exists⁸. However, it seems that thermal fluctuations are suppressed for an array — an observation that calls for further investigation⁹.

We have found that superfluid ^4He in an array of small apertures behaves quantum coherently, oscillating at the Josephson frequency. Because these oscillations appear in ^4He at a temperature 2,000 times higher than

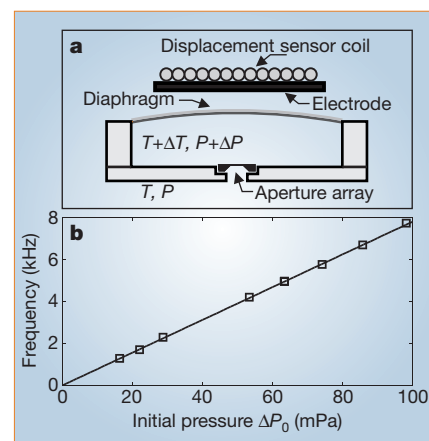


Figure 1 Quantum oscillations in ^4He . **a**, Experimental cell (see supplementary information for details). **b**, Whistle frequency plotted against the initial pressure, $\Delta P_0 = \rho\Delta\mu_0/m_4$. Temperature is in the range where, if T_λ is the superfluid transition temperature, $T_\lambda - T$ is 1.7–2.9 mK. A fit (solid line) to the data gives a slope of 78 Hz mPa^{-1} , with a systematic uncertainty of 20% arising from our pressure calibration. This agrees with the Josephson frequency relation $f_j = \Delta\mu/h$ value of 68.7 Hz mPa^{-1} . The oscillation is still present down to at least 150 mK below T_λ , where the healing length is much smaller than the aperture diameter and $I(\Delta\phi)$ is linear. The oscillation is presumably due to periodic 2π phase slips.

in superfluid ^3He , it may be possible to build sensitive rotation sensors using much simpler technology than previously believed^{10–13}. This could find application in rotational seismology, geodesy and tests of general relativity.

E. Hoskinson, R. E. Packard,

Thomas M. Haard

Physics Department, University of California, Berkeley, California 94720, USA

e-mail: packard@socrates.berkeley.edu

1. Van Duzer, T. & Turner, C. W. *Superconductive Devices and Circuits* 2nd edn (Prentice-Hall, Upper Saddle River, 1999).
2. Pereversev, S. V., Loshak, A., Backhaus, S., Davis, J. C. & Packard, R. E. *Nature* **388**, 449–451 (1997).
3. Davis, J. C. & Packard, R. E. *Rev. Mod. Phys.* **74**, 741–773 (2002).
4. Tilley, D. R. & Tilley, J. in *Superfluidity and Superconductivity* 3rd edn, Ch. 7 (Institute of Physics, Bristol and Philadelphia, 1990).
5. Sukhatme, K., Mukharsky, Y., Chui, T. & Pearson, D. *Nature* **411**, 280–283 (2001).
6. Avenel, O. & Varoquaux, E. *Phys. Rev. Lett.* **55**, 2704–2707 (1985).
7. Packard, R. E. & Vitale, S. *Phys. Rev. B* **45**, 2512–2515 (1992).
8. Backhaus, S. & Packard, R. E. *Phys. Rev. Lett.* **81**, 1893–1896 (1998).
9. Chui, T., Holmes, W. & Penanen, K. *Phys. Rev. Lett.* **90**, 085301 (2003).
10. Packard, R. E. & Vitale, S. *Phys. Rev. B* **46**, 3540–3549 (1992).
11. Avenel, O., Hakonen, P. & Varoquaux, E. *Phys. Rev. Lett.* **78**, 3602–3605 (1997).
12. Schwab, K., Bruckner, N. & Packard, R. E. *Nature* **386**, 585–587 (1997).
13. Simmonds, R. W., Marchenkov, A., Hoskinson, E., Davis, J. C. & Packard, R. E. *Nature* **412**, 55–58 (2001).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

brieF communications arising online

www.nature.com/bca

Marine ecology: Different measures of biodiversity

J. R. Dolan (doi:10.1038/nature03320)

Reply: X. Irigoien, J. Huisman, R. P. Harris

(doi:10.1038/nature03321)

Marine ecology

Different measures of biodiversity

Arising from: X. Irigoien, J. Huisman & R. P. Harris *Nature* **429**, 863–867 (2004)

In their analysis of global trends of diversity in marine plankton, Irigoien *et al.*¹ find that taxonomic diversity of zooplankton (consumer) is a unimodal function of community biomass that is unrelated to phytoplankton (producer) diversity. Their results are unexpected because in terrestrial organisms primary-producer diversity is a good predictor of consumer diversity². I contend that this apparent uncoupling of producer and consumer diversity in marine plankton is likely to be an artefact due to the authors' use of different measures of diversity for producers and for consumers.

The relationships described by Irigoien *et al.*¹ between phytoplankton and zooplankton diversity on a global scale were based on analysis of 100-ml samples examined in an inverted microscope. This method is commonly used for species identification and abundance estimates of phytoplankton, although it underestimates the diversity of small (< 10 µm) species³. For zooplankton, the same method can furnish estimates only of the diversity of easily distinguishable morphotypes, not of species.

The sample material examined by Irigoien *et al.*¹ had been fixed with either Lugol's solution or glutaraldehyde, which in the case of microzooplankton allows reliable species identification in only a very small fraction of cells — namely those with a distinct gross morphology, such as tintinnid ciliates, radiolarians and foraminifera that have 'shells' or skeletons: these identifiable forms generally represent some 5–10% of the community. For most ciliates and dinoflagellates, identification is, at best, limited to the genus⁴. This fixation method is useful for biomass estimations, enabling ciliates and dinoflagellate cells to be placed in size–shape categories; however, these may or may not have ecological relevance and are unlikely to be of any taxonomic validity.

Moreover, Irigoien *et al.*¹ analyse very different population sizes, which might also bias their diversity estimates. Producer and consumer diversity were examined in the same sample, which had relatively few consumers: in 100 ml of surface water, phytoplankton cells number 10³–10⁴, compared with 1–10² microzooplankton cells. This explains in part why microzooplankton diversity was positively related to biomass or numbers of individuals, and also why a total of only 43 'species' of microzooplankton were encountered in this global study. This number is low; large volumes from a single sampling site can yield 39 species of a numerically minor component of tintinnid ciliate microzooplankton⁵.

I therefore propose that the data presented by Irigoien *et al.* show a weak relation not

between taxonomic diversity of consumers and producers but between the size diversity of consumers and taxonomic diversity of producers. Although there could be a correspondence of diversity in producers and consumers if appropriate estimation measures were to be used, this remains to be investigated.

John R. Dolan

Marine Microbial Ecology Group, Laboratoire de Océanographie de Villefranche, Université de Paris VI et CNRS UMR 7093 Station Zoologique, BP 28, 06230 Villefranche-sur-Mer, France
e-mail: dolan@obs-vlfr.fr
doi:10.1038/nature03320

1. Irigoien, X., Huisman, J. & Harris, R. P. *Nature* **429**, 863–867 (2004).
2. Morin, P. J. & Fox, J. W. *Nature* **429**, 813–814 (2004).
3. Hasle, G. R. in *Phytoplankton Manual* (ed. Sournia, A.) 125–128 (UNESCO, Paris, 1978).
4. Gifford, D. J. & Caron, D. A. in *Zooplankton Methodology Manual* (eds Harris, R. P., Weibe, P. H., Lenz, J., Skjoldal, H. R. & Huntley, M.) 193–221 (Academic Press, London, 2000).
5. Cariou, J.-B., Dolan, J. R. & Dallot, S. J. *Plank. Res.* **21**, 1065–1075 (1999).

Irigoien et al. reply — We agree with Dolan's concern that simple morphological traits reveal only part of the zooplankton biodiversity¹. We argue, however, that this is unlikely to affect our findings².

Dolan doubts the lack of relation between phytoplankton diversity and zooplankton diversity in the world's oceans². He argues that we compared apples with oranges, because our method to identify microzooplankton using simple morphological traits was limited to the genus level and hid microzooplankton diversity at the species level.

This argument applies not only to microzooplankton, but to many other taxonomic groups as well. Ecologists are often confronted with problems of species identity when estimating the diversity of organisms within poorly known groups. This is as true for studies of insects in tropical forests as it is for those of plankton in the oceans. Genetic studies of the taxonomy of small cyanobacteria³, of morphologically well established species of diatoms⁴, and of conspicuous organisms as large as the copepods⁵, have all revealed cryptic species. We therefore fully agree with Dolan that our simple methods will have underestimated the diversity at the species level.

However, our metrics would have been biased only if we had compared diversity data obtained by different methods (for example, morphological diversity for phytoplankton compared with genetic diversity for zooplankton). We disagree with Dolan's critique that we used two different measures of diversity. Instead, to avoid bias, our biodiversity estimates of phytoplankton and zooplankton were both based on

exactly the same methodology, using only simple morphological traits.

We suggest that the lack of a positive correlation between plant diversity and herbivore diversity in the oceans is real, and can be explained by the broad diet of most marine herbivores. In contrast to many specialized terrestrial herbivores (many insects, for example), most zooplankton species do not rely on their food having a specific taxonomic composition. In terrestrial systems with a tight association between species (such as insects and their host plants), even taxonomic lumping of the sort criticized by Dolan is likely to reveal a positive correlation between plant diversity and herbivore diversity at the level of genus or family. Indeed, a seminal work on plant–insect coevolution in terrestrial systems focuses on the family and genus level⁶.

Perhaps, then, the observed lack of a relationship was an artefact because we did not distinguish enough morphospecies among the microzooplankton? In which case, the alternative hypothesis of an increase in microzooplankton diversity with phytoplankton diversity would predict a decrease in the variance of microzooplankton diversity. However, we found that the variance in microzooplankton diversity was not related to phytoplankton diversity either. Inspection of our data reveals that the lack of a relationship may not be due to an underestimation of microzooplankton diversity, but rather to a relatively high microzooplankton diversity in oligotrophic regions with relatively low phytoplankton diversity. Likewise, a high copepod diversity has been found in oligotrophic areas of the Atlantic Ocean⁷.

The advance of molecular genetic techniques in oceanographic research may resolve the issue. Information on the genetic diversity of marine phytoplankton and zooplankton could confirm our findings⁸ and spur further research⁸ in this direction.

Xabier Irigoien*, **Jef Huisman†**,
Roger P. Harris‡

**AZTI, Arantzeta eta Elikagintzarako Institutu Teknologikoa, Herrera Kaia Portualdea, 20110 Pasaia, Spain*
e-mail: xirigoien@pas.azti.es

†Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands

‡Plymouth Marine Laboratory, Prospect Place, Plymouth PL1 3DH, UK
doi:10.1038/nature03321

1. Dolan, J. *Nature* doi:10.1038/nature03320 (2005).
2. Irigoien, X., Huisman, J. & Harris, R. P. *Nature* **429**, 863–867 (2004).
3. Janse, L., Meima, M., Kardinaal, W. E. A. & Zwart, G. *Appl. Environ. Microbiol.* **69**, 6634–6643 (2003).
4. Orsini, L., Procaccini, G., Sarno, D. & Montresor, M. *Mar. Ecol. Prog. Ser.* **271**, 87–98 (2004).
5. Goetze, E. *Proc. R. Soc. Lond. B* **270**, 2321–2331 (2003).
6. Ehrlich, P. R. & Raven, P. H. *Evolution* **18**, 586–608 (1964).
7. Woodd-Walker, R. S., Ward, P. & Clarke, A. *Mar. Ecol. Prog. Ser.* **236**, 189–203 (2002).
8. Census of Marine Life (www.coml.org).

Recognition of transmembrane helices by the endoplasmic reticulum translocon

Tara Hessa¹, Hyun Kim¹, Karl Bihlmaier^{1*}, Carolina Lundin¹, Jorrit Boekel¹, Helena Andersson³, IngMarie Nilsson¹, Stephen H. White² & Gunnar von Heijne¹

¹Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden

²Department of Physiology and Biophysics and the Program in Macromolecular Structure, University of California at Irvine, Irvine, California 92697-4560, USA

³Karolinska Institutet, Department of Bioscience at NOVUM, SE-141 57 Huddinge, Sweden

* Present address: Institute of Biochemistry, ETH Zurich, CH-8093 Zurich, Switzerland

Membrane proteins depend on complex translocation machineries for insertion into target membranes. Although it has long been known that an abundance of nonpolar residues in transmembrane helices is the principal criterion for membrane insertion, the specific sequence-coding for transmembrane helices has not been identified. By challenging the endoplasmic reticulum Sec61 translocon with an extensive set of designed polypeptide segments, we have determined the basic features of this code, including a 'biological' hydrophobicity scale. We find that membrane insertion depends strongly on the position of polar residues within transmembrane segments, adding a new dimension to the problem of predicting transmembrane helices from amino acid sequences. Our results indicate that direct protein–lipid interactions are critical during translocon-mediated membrane insertion.

Integral membrane proteins account for 20–30% of all genes in both prokaryotic and eukaryotic organisms¹ and are involved in a host of biological functions such as signal transduction, solute and macromolecular transport across membranes, cell–cell interactions and nerve conduction. The great majority of integral membrane proteins belong to the helix-bundle class², that is, their basic architecture is one of tightly packed transmembrane (TM) α -helices.

In eukaryotic cells, most helix-bundle membrane proteins insert co-translationally and fold in the endoplasmic reticulum (ER) membrane³. Insertion is mediated by the Sec61 translocon, a hetero-oligomeric protein-conducting channel⁴. Individual TM helices seem to follow an ordered insertion pathway, in which they pass from the tunnel in the large ribosomal subunit into the Sec61 translocon channel and then exit the channel laterally into the surrounding lipid^{5,6}. The recently determined X-ray structure of an archaeal SecY translocon (homologous to the eukaryotic Sec61 translocon) suggests that segments in a translocating polypeptide can move laterally into the ER membrane through a gate located between helices 2b–3 and 7–8 in the Sec61 α -subunit to form TM helices⁷.

A fundamental question concerns the design of the 'code' that determines whether or not a polypeptide segment is recognized by the translocon for integration into the ER membrane⁸. Here, we approach this question by comparing extensive data on TM helix integration obtained from quantitative studies of the translocon-mediated pathway with data obtained from biophysical studies of well defined model systems^{9,10} and from statistical analyses of membrane protein structures^{11,12}. Overall, our results strongly indicate that direct interactions between the nascent polypeptide in transit through the ER membrane and the lipid bilayer surrounding the Sec61 translocon are critical in the recognition of TM helices.

of hydrophobicity, we challenged the Sec61 translocon with a large set of systematically designed TM sequences and measured the efficiency of membrane integration for each one. To the extent that quantitative data generated in this way can be decomposed into contributions from individual residues, the output from the biological system can be directly compared with biophysical measurements.

Precise quantification of the biological process is essential for such comparisons. We use an *in vitro* assay¹³ for quantifying the efficiency of membrane integration of designed TM segments into dog pancreas rough microsomes (RMs). Briefly, a segment to be tested (H-segment) is engineered into the luminal P2 domain of the integral membrane protein leader peptidase (Lep), where it is flanked by two acceptor sites for N-linked glycosylation (Fig. 1a). The degree of membrane integration of the H-segment is quantified from SDS–PAGE gels (Fig. 1b) by measuring the fraction of singly (f_{1g}) versus singly plus doubly (f_{2g}) glycosylated Lep molecules, $p = f_{1g}/(f_{1g} + f_{2g})$. The data can also be expressed as an apparent equilibrium constant, $K_{app} = f_{1g}/f_{2g}$, between the membrane integrated and non-integrated forms. Although this expression is not meant to imply that the membrane integration process reflects a true thermodynamic equilibrium (but this does seem to be a good first approximation, see below), it has the advantage that the results can be converted to apparent free energies, $\Delta G_{app} = -RT \ln K_{app}$, for direct comparison with biophysical data.

We have studied H-segments with the general design GGPG-X₁₉-GPGG, in which the flanking tetrapeptides are included to 'insulate' the central 19-residue stretch from the surrounding sequence. The flanking tetrapeptides were chosen to be indifferent with respect to charge and other specific interactions that might occur in the membrane and/or translocon, but to have a low probability of secondary structure formation.

Connecting biological and biophysical principles

The high content of nonpolar residues in TM segments is a clear indication that hydrophobicity is an important component of the TM-sequence code read by the translocon. To assess the importance

A biological hydrophobicity scale

The Wimley–White water/octanol free energy scale^{10,14} predicts that H-segments composed of n leucine and $(19 - n)$ alanine residues will be stably inserted across the membrane for $n \approx 5$ – 6 ¹⁵. We thus

began by testing H-segments of the design GGPG-(L_nA_{19-n})-GPGG with $n = 0-7$. Results obtained when 2–4 Leu residues are incorporated in various positions along the sequence are shown in Fig. 1c. There is an overall linear relationship between n and ΔG_{app} , a simple outcome consistent with energy-additivity. Furthermore, the probability of insertion, p , conforms to a Boltzmann distribution (Fig. 1d). The Boltzmann distribution shows that translocation-mediated insertion has the appearance of an equilibrium process, supporting the equilibrium approximation behind the calculation of ΔG_{app} . Additionally, Fig. 1c shows that for a given n , there is a variation in ΔG_{app} of $\pm 0.2-0.3$ kcal mol⁻¹ around the mean when the positions of the Leu residues are changed. Thus, total hydrophobicity is not the sole determinant of membrane insertion; residue position also has an important influence.

Ignoring for the moment the positional variation, the data in Fig. 1c can be fitted by the expression:

$$\begin{aligned}\Delta G_{app} &= -0.66n + 2.14 = n\Delta G_{app}^{Leu} + (19-n)\Delta G_{app}^{Ala} + \Delta G_{app}^{flank} \\ &= n(\Delta G_{app}^{Leu} - \Delta G_{app}^{Ala}) + 19\Delta G_{app}^{Ala} + \Delta G_{app}^{flank}\end{aligned}\quad (1)$$

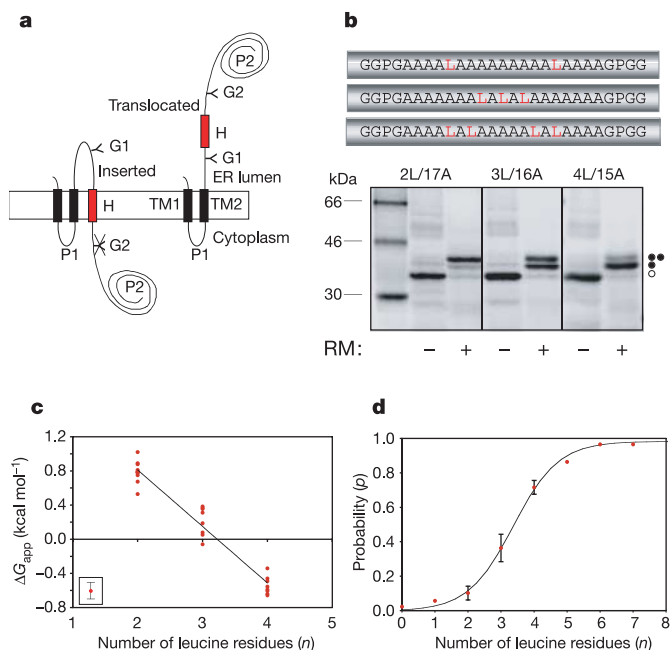


Figure 1 Integration of H-segments into the microsomal membrane. **a**, Wild-type Lep has two N-terminal TM segments (TM1 and TM2) and a large luminal domain (P2). H-segments were inserted between residues 226 and 253 in the P2-domain. Glycosylation acceptor sites (G1 and G2) were placed in positions 96–98 and 258–260, flanking the H-segment. For H-segments that integrate into the membrane, only the G1 site is glycosylated (left), whereas both the G1 and G2 sites are glycosylated for H-segments that do not integrate in the membrane (right). **b**, Membrane integration of H-segments with the Leu/Ala composition 2L/17A, 3L/16A and 4L/15A. Plasmids encoding the Lep/H-segment constructs were transcribed and translated *in vitro* in the presence (+) and absence (–) of dog pancreas rough microsomes (RM). Bands of unglycosylated protein are indicated by a white dot; singly and doubly glycosylated proteins are indicated by one and two black dots, respectively. **c**, ΔG_{app} values for H-segments with 2–4 Leu residues. The average standard deviation in the individual ΔG_{app} determinations is shown in the boxed insert. Individual points for a given n show ΔG_{app} values obtained when the position of Leu is changed. **d**, Mean probability of insertion (p) for H-segments with $n = 0-7$ Leu residues. The curve shown represents the best-fit Boltzmann distribution. For $n = 2-4$, mean values and standard deviations were calculated by averaging the data in **c**. For $n = 0, 1, 5-7$, only single H-segments with the following compositions were used (flanked by GGPG...GPGG in all cases): (A)₁₉, (A)₉L(A)₉, (A)₄LALAALA(LA)₄, (A)₄(LA)₅L(A)₄, ALAALAALAALAALA.

where the last term accounts for the contribution to ΔG_{app} from the two tetrapeptides flanking the central 19-residue segment. The average $\Delta\Delta G_{app}^{Ala \rightarrow Leu}$ for an Ala $\rightarrow$ Leu replacement is thus -0.7 kcal mol⁻¹. On the assumption that $\Delta G_{app}^{flank} = 0$ (see below), we can further calculate individual ΔG_{app}^{aa} values: $\Delta G_{app}^{Ala} = 0.1$ kcal mol⁻¹ and $\Delta G_{app}^{Leu} = -0.6$ kcal mol⁻¹.

We next determined ΔG_{app}^{aa} values for the remaining 18 naturally occurring amino acids when placed in the middle of the 19-residue stretch. The quantification is maximally sensitive for H-segments

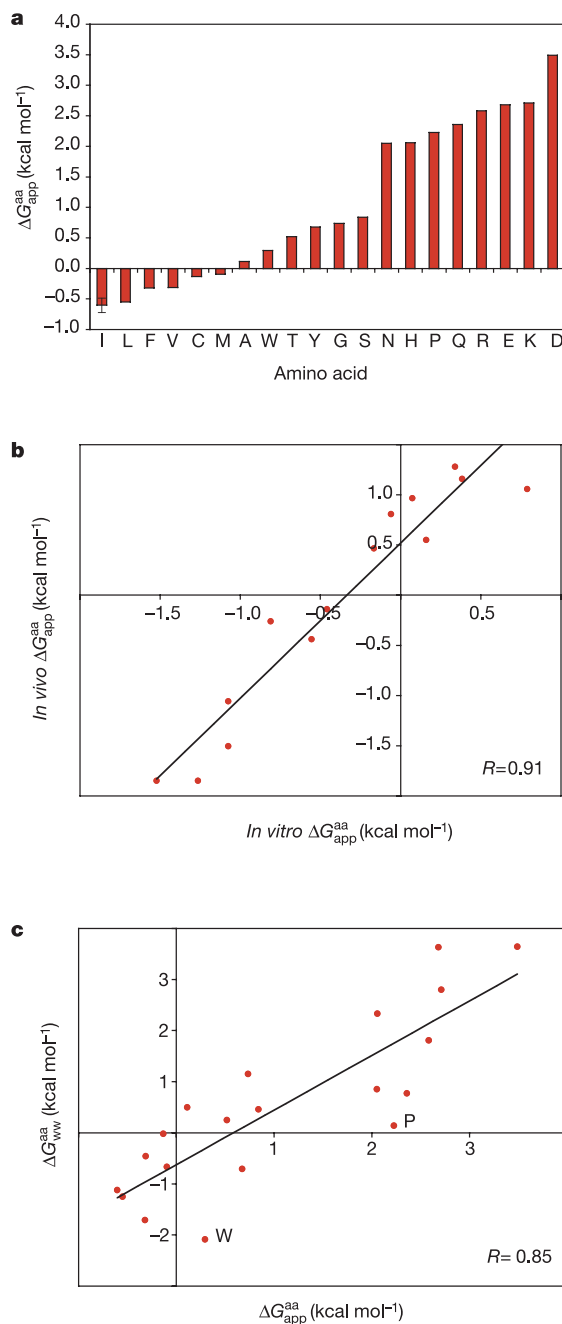


Figure 2 Biological and biophysical ΔG^{aa} scales. **a**, ΔG_{app}^{aa} scale derived from H-segments with the indicated amino acid placed in the middle of the 19-residue hydrophobic stretch (see Supplementary Information S1 for details). The bar indicates the standard deviation in the determination of ΔG_{app}^{Leu} ; the standard deviation for all other amino acids is similar. **b**, Correlation between ΔG_{app}^{aa} values measured *in vivo* and *in vitro*. See Supplementary Information S3 for details on the H-segments used. **c**, Correlation between the ΔG_{app}^{aa} scale and the Wimley–White water/octanol free energy scale¹⁴ (ΔG_{WW}^{aa}).

with ΔG_{app} values close to zero ($p \approx 0.5$ in Fig. 1d); therefore, for each kind of residue we balanced the contribution from the central residue by varying the number of Leu residues until an H-segment with ΔG_{app} in the range -1.2 to 1.2 kcal mol $^{-1}$ was found. Based on these data, a full $\Delta G_{\text{app}}^{\text{aa}}$ scale was calculated using a stepwise procedure (described in Supplementary Information S1). As seen in Fig. 2a, Ile, Leu, Phe and Val have $\Delta G_{\text{app}}^{\text{aa}} < 0$ and thus promote membrane insertion, Cys, Met and Ala have $\Delta G_{\text{app}}^{\text{aa}} \approx 0$, and all polar and charged residues have $\Delta G_{\text{app}}^{\text{aa}} > 0$. A charged residue can be tolerated in the centre of a TM segment if its unfavourable cost is offset by a sufficiently large number of nonpolar residues, as expected for thermodynamic equilibrium. Related H-segments in which either one or two Ala residues have been changed to Gly, Ser, Trp or Tyr all have $\Delta \Delta G_{\text{app}}^{\text{Ala} \rightarrow \text{X}} \approx 0.5 \Delta \Delta G_{\text{app}}^{\text{Ala} \rightarrow 2\text{X}}$ (see Supplementary Information S2), suggesting that the $\Delta G_{\text{app}}^{\text{aa}}$ scale is approximately additive. To ensure that the *in vitro* results are relevant to the *in vivo* situation, selected constructs were also expressed *in vivo* in BHK cells. On average, the ΔG_{app} values changed by only 0.5 kcal mol $^{-1}$, corresponding to an increase in the individual $\Delta G_{\text{app}}^{\text{aa}}$ values by 0.03 kcal mol $^{-1}$ compared to the *in vitro* results (Fig. 2b).

The 'biological' $\Delta G_{\text{app}}^{\text{aa}}$ scale clearly has much in common with hydrophobicity scales derived from biophysical measurements^{16,17}

and with statistical analyses of the lipid-exposed parts of high-resolution membrane protein structures^{11,12}. The correlation between the $\Delta G_{\text{app}}^{\text{aa}}$ scale and the Wimley–White water/octanol free energy scale¹⁴ is shown in Fig. 2c. Considering the complexity of the biological system, the two scales correlate surprisingly well: the linear fit has a slope of 1.1 and the origins of the two scales coincide within 0.5 kcal mol $^{-1}$. The only clear amino acid outliers are Trp (W) and Pro (P), which are both more hydrophobic on the Wimley–White scale. The correlation between the two scales (and between the $\Delta G_{\text{app}}^{\text{aa}}$ scale and other biophysical scales, data not shown) immediately suggests that the recognition of TM segments by the translocon involves direct interaction between the TM segment and the surrounding lipid¹⁸.

Can the contribution from the flanking tetrapeptides ($\Delta G_{\text{app}}^{\text{flank}}$) significantly affect these conclusions? To address this question, we first lengthened the flanking stretches of Gly residues stepwise from GGPG-X₁₉-GP GG to GGGGGGPG-X₁₉-GGGGGGG for an H-segment with 3 Leu and 16 Ala residues (3L/16A). For this series, the ΔG_{app} values vary by no more than ± 0.2 kcal mol $^{-1}$ (see Supplementary Information S4), demonstrating that residues in the Lep P2 domain outside the H-segment have little influence on the results. We also replaced the GGPG-X₁₉-GP GG flanks by NNPN-X₁₉-NPNN and tested H-segments containing three or

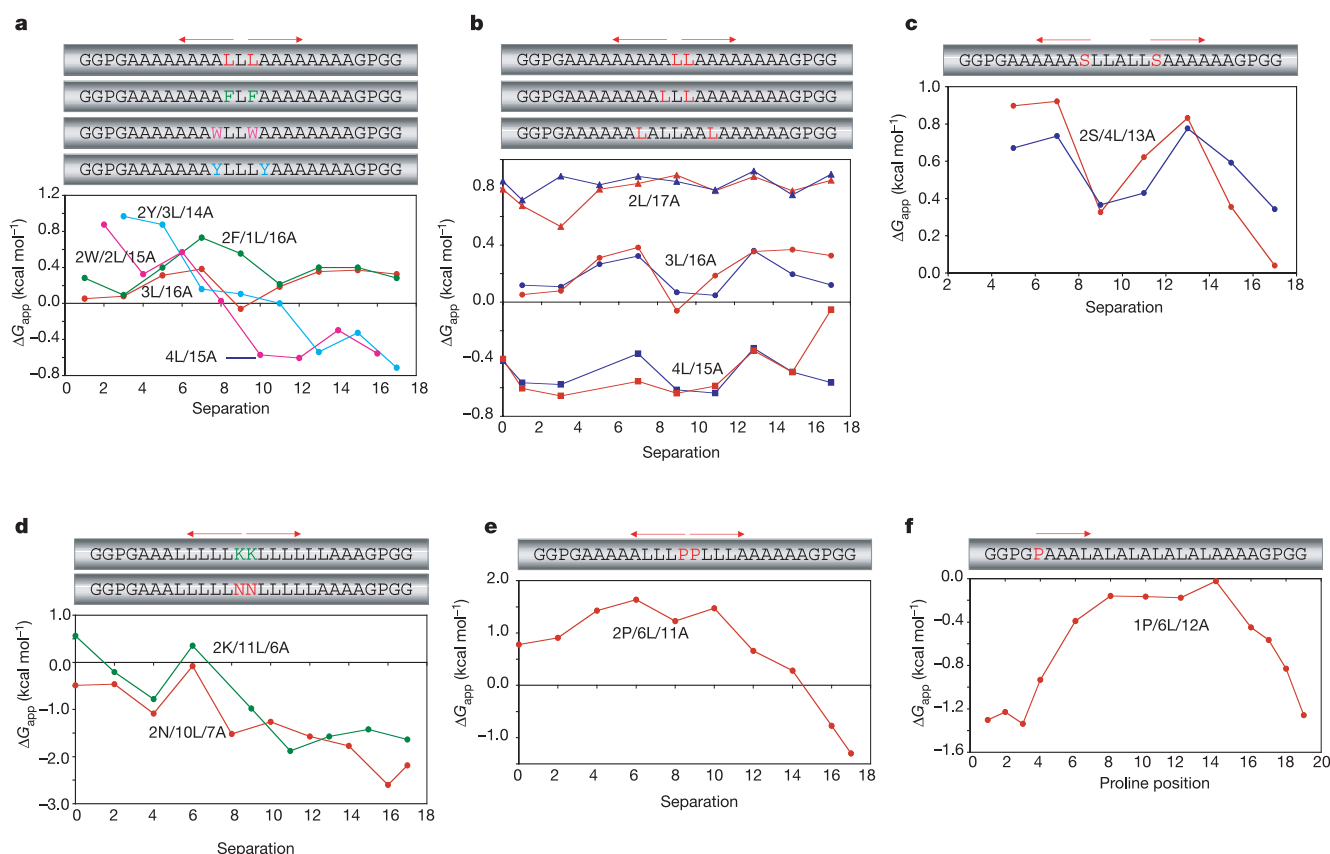


Figure 3 Positional dependencies in ΔG_{app} . **a**, Symmetrical H-segment scans with pairs of Leu (red), Phe (green), Trp (pink) or Tyr (light blue) residues. The Leu scan is based on symmetrical 3L/16A H-segments with a Leu-Leu separation of one residue (sequence shown at the top; the two red Leu residues are moved symmetrically outwards) up to a separation of 17 residues. For the Phe scan, the composition of the central 19-residues of the H-segments is 2F/1L/16A, for the Trp scan it is 2W/2L/15A, and for the Tyr scan it is 2Y/3L/14A. The ΔG_{app} value for the 4L/15A H-segment GGPGAAALAAAAAALAAAGP GG is also shown (dark blue). **b**, Red lines show ΔG_{app} values for symmetrical scans of 2L/17A (triangles), 3L/16A (circles), and 4L/15A (squares) H-segments. Blue lines show the ΔG_{app} values predicted from least-square

linear fits between the observed data and the hydrophobic moment ($\delta = 100^\circ$) of the central 19-residue segments (see Methods). The predicted values all have the form $\Delta G_{\text{app}} = 0.17\mu + h_i$, where μ is the hydrophobic moment and h_i is the extrapolated ΔG_{app} value for an H-segment of the given composition with $\mu = 0$. **c**, Same as **b** but for a symmetrical scan with pairs of Ser residues in H-segments with the composition 2S/4L/13A. **d**, Symmetrical scans with pairs of Lys (green) and Asn (red) residues in H-segments with the overall compositions 2K/11L/6A and 2N/10L/7A. **e**, Symmetrical scan with a pair of Pro residues in H-segments with the composition 2P/6L/11A. **f**, Scan with a single Pro residue in H-segments with the composition 1P/6L/12A. The position of the Pro residue in the 19-residue central hydrophobic stretch is shown on the x-axis.

four Leu residues. On average, ΔG_{app} increases by only $0.5 \text{ kcal mol}^{-1}$ when the six Gly residues in the flanking tetrapeptides are replaced by Asn (see Supplementary Information S4), suggesting that ΔG_{app}^{flank} is indeed close to zero for the H-segments used to establish the ΔG_{app} scale and that the true zero-point on the ΔG_{app}^{aa} scale is close to ΔG_{app}^{Ala} . Interestingly, studies of model peptides composed of 18–23 Ala residues have concluded that poly-Ala segments are at the threshold for transmembrane integration into pure lipid bilayers^{9,19–21}.

Effects of amino acid position on TM helix insertion

The ΔG_{app}^{aa} scale is strictly valid only for residues placed in the middle of the H-segment. To examine how ΔG_{app}^{aa} varies with residue position, we performed scans in which a pair of identical amino acid residues were moved symmetrically from the centre of the H-segment towards its amino and carboxy termini. Results for symmetrical scans with pairs of Leu, Phe, Trp and Tyr residues are shown in Fig. 3a. Leu and Phe behave similarly and show a small but significant position-specific variation of $\pm 0.2 \text{ kcal mol}^{-1}$ around the mean. Trp and Tyr show markedly different behaviour: they strongly reduce membrane insertion when placed centrally, but become much less unfavourable as they are moved apart. Indeed, Trp is as favourable as Leu when placed in the outermost positions (see the construct 4L/15A in Fig. 3a). In membrane proteins, Trp and Tyr (in contrast to Phe and Leu) are over-represented near the ends of TM helices^{11,12,22} and are known to interact preferentially with the headgroup region of model phospholipid bilayers^{23–25}. The results presented here provide further support for the idea that protein–lipid interactions are central to the recognition of TM helices by the translocon.

Can we rationalize the variation in ΔG_{app} around the mean in the symmetrical scans? An obvious possibility is that these variations reflect the amphiphilicity of the H-segment. We used the ΔG_{app}^{aa} scale to compute the helical hydrophobic moment²⁶ for the central 19-residue segment in symmetrical scans with Leu and Ser residues, and then calculated the least-squares linear fit between experimentally-derived ΔG_{app} values and the hydrophobic moment. The ΔG_{app} values can be quite well reproduced by this procedure (Fig. 3b, c). The least amphiphilic H-segments (lowest hydrophobic moment) insert most efficiently into the membrane, and vice versa. The picture is similar for symmetrical scans with Lys and Asn residues, (Fig. 3d). Terminal Asn and Lys residues are much less detrimental to membrane insertion than centrally located ones, and there is a very strong amphiphilicity effect for H-segments with the two polar residues separated by six residues (that is, with the polar residues on the same face of an α -helix), which have greatly increased ΔG_{app} values. An interesting possibility suggested by these findings is that the H-segment forms an α -helix oriented so that the less hydrophobic face is in contact with the translocon protein and the more hydrophobic face is in contact with lipid during the membrane insertion process, consistent with crosslinking data^{27,28}.

To test further the importance of helical structure, scans with a single Pro or a pair of Pro residues were also done. In contrast with all the other residues tested, a pair of Pro residues maximally reduces the efficiency of membrane insertion at intermediate separations (4–10 residues; Fig. 3e). The positional dependence in ΔG_{app} is very strong, with a 1 kcal mol^{-1} difference between a centrally located Pro-Pro pair and a pair separated by 4–10 residues, and an even greater ΔG_{app} difference between centrally and terminally located Pro pairs. Furthermore, the scan with a single Pro residue (Fig. 3f) reveals that H-segments with a Pro residue in any of the three N-terminal positions of the hydrophobic stretch integrate much more efficiently than H-segments with Pro in the three C-terminal positions. In globular proteins, Pro is frequently found in the three N-terminal positions of α -helices, but is almost completely absent from the central and C-terminal parts^{29,30}. Like-

wise, Pro is well tolerated near the N termini of TM helices in bacteriorhodopsin³¹. We conclude that helix formation is an important part of the membrane insertion process.

Can the biological ΔG_{app}^{aa} scale presented here be used to calculate the membrane insertion efficiency of natural polypeptide segments? It could in principle, but the scale is still too crude to be used for precise calculations of this kind because of the positional dependence of the ΔG_{app}^{aa} values, the contribution of charged flanking residues to ΔG_{app} (ref. 32), and the possible role of helix–helix interactions during insertion of multi-spanning membrane proteins³³; further work will be necessary to quantify these effects fully.

In summary, the similarity between the biological ΔG_{app}^{aa} scale and both biophysical and statistical hydrophobicity scales, together with the positional preferences seen for residues such as Trp, Tyr, Ser, Asn, Lys and Pro, strongly suggest that direct protein–lipid interactions are essential for the recognition of TM helices by the translocon, and support models based on a partitioning of the TM helices between the Sec61 translocon and the surrounding lipid⁵. Exactly how this is managed is not apparent from the archaeal translocon structure, which is observed in a closed state without a translocating TM segment. Presumably, the open state is a highly dynamic one that permits rapid sampling of the translocon–bilayer interface by the translocating polypeptide. □

Methods

Enzymes and chemicals

All enzymes, the plasmid pGEM1, dithiothreitol (DTT) and the TnT coupled transcription/translation system were from Promega. ³⁵S-Met, ribonucleotides, deoxyribonucleotides, and dideoxyribonucleotides were from Amersham-Pharmacia. Oligonucleotides were obtained from Cybergene and MWG Biotech AG.

DNA manipulations

For expression of H-segment-containing Lep constructs from the pGEM1 plasmid, the 5' end of the *lep* gene from *Escherichia coli* was modified by the introduction of an *Xba*I site and by changing the context 5' of the initiator ATG codon to a Kozak consensus sequence³⁴. Site-directed mutagenesis was used to introduce acceptor sites for N-linked glycosylation in positions 96–98 (Asn-Ser-Thr) and 258–260 (Asn-Ala-Thr).

Oligonucleotides encoding the different H-segments were introduced between a *Spe*I cleavage site in codons 226–227 and the *Kpn*I cleavage site in codon 253 of the *lep* gene¹³. H-segments were constructed using two or three double-stranded oligonucleotides (18–48 nucleotides long) with overlapping overhangs at the ends. Pairs of complementary oligonucleotides were first annealed at 85 °C for 10 min followed by slow cooling to 30 °C, after which the two or three annealed double-stranded oligos were mixed, incubated at 65 °C for 5 min, cooled slowly to room temperature and ligated into the vector. All H-segment inserts were confirmed by sequencing of plasmid DNA.

Expression in vitro

Constructs in pGEM1 were transcribed and translated in the Gold TnT Express 96 or TnT Quick systems (Promega). 1 μ g DNA template, 1 μ l ³⁵S-Met (5 μ Ci) and 1 μ l microsomes (a gift from M. Sakaguchi) were added at the start of the reaction, and samples were incubated for 90 min at 30 °C. Translation products were analysed by SDS–PAGE and gels were quantified using a Fuji FLA-3000 phosphorimager and Image Reader 8.1j software. The membrane-insertion probability of a given H-segment was calculated as the quotient between the intensity of the singly glycosylated band divided by the summed intensities of the singly glycosylated and doubly glycosylated bands. Mean values from at least four independent experiments were used in the derivation of the ΔG_{app}^{aa} scale, and from at least two independent experiments for the other results reported. On average, glycosylation levels vary by no more than $\pm 2\%$ between repeat experiments, corresponding to a standard deviation of $\pm 0.1 \text{ kcal mol}^{-1}$ in the ΔG_{app} values.

Expression in vivo

Protein synthesis in BHK cells using the Semliki forest virus (SFV) expression system was carried out as described previously^{35,36}. Briefly, Lep constructs under the SP6 promoter in the SFV vector were linearized for *in vitro* transcription. The resulting RNA was used to transfect BHK cells by electroporation. Six hours after electroporation, cells were starved of Met for 30 min, then labelled with ³⁵S-Met for 15 min. Cells were solubilized in lysis buffer containing 1% nonidet P-40 and protease inhibitors, immunoprecipitated using a Lep antiserum and analysed by SDS–PAGE. Quantifications were carried out on a phosphorimager as described above.

Hydrophobic moment calculation

Hydrophobic moments (μ) were computed using MPEx (<http://blanco.biomol.uci.edu/mpex/>) and the ΔG_{app}^{aa} scale. Linear fits ($\Delta G_{app} = a\mu + b$) were calculated between μ and the ΔG_{app} values for a given symmetrical pair scan (*i*), excluding the data point representing the maximum separation (17 residues) between the scanned residues. For the

2L/17A series, the point representing a separation of 3 residues was also excluded. Finally, the a_i values were averaged for the pair scans with 2L/17A, 3L/16A, 4L/15A and 2S/4L/13A, yielding the mean value $a = 0.17$ (standard deviation, ± 0.04).

Received 31 August; accepted 16 November 2004; doi:10.1038/nature03216.

1. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
2. von Heijne, G. Recent advances in the understanding of membrane protein assembly and structure. *Q. Rev. Biophys.* **32**, 285–307 (2000).
3. von Heijne, G. Membrane protein assembly *in vivo*. *Adv. Protein Chem.* **63**, 1–18 (2003).
4. Snapp, E., Reinhard, G., Bogert, B., Lippincott-Schwartz, J. & Hegde, R. The organization of engaged and quiescent translocons in the endoplasmic reticulum of mammalian cells. *J. Cell Biol.* **164**, 997–1007 (2004).
5. Rapoport, T. A., Goder, V., Heinrich, S. U. & Matlack, K. E. Membrane-protein integration and the role of the translocation channel. *Trends Cell Biol.* **14**, 568–575 (2004).
6. Alder, N. N. & Johnson, A. E. Cotranslational membrane protein biogenesis at the endoplasmic reticulum. *J. Biol. Chem.* **279**, 22787–22790 (2004).
7. van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427**, 36–44 (2004).
8. Woolhead, C. A., McCormick, P. J. & Johnson, A. E. Nascent membrane and secretory proteins differ in FRET-detected folding. *Cell* **116**, 725–736 (2004).
9. de Planque, M. R. R. & Killian, J. A. Protein-lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring. *Mol. Membr. Biol.* **20**, 271–284 (2003).
10. White, S. H. & Wimley, W. C. Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365 (1999).
11. Ulmschneider, M. B. & Sansom, M. S. P. Amino acid distributions in integral membrane protein structures. *Biochim. Biophys. Acta* **1512**, 1–14 (2001).
12. Beuming, T. & Weinstein, H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics* **20**, 1822–1835 (2004).
13. Sääf, A., Wallin, E. & von Heijne, G. Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *Eur. J. Biochem.* **251**, 821–829 (1998).
14. Wimley, W. C., Creamer, T. P. & White, S. H. Solvation energies of amino acid sidechains and backbone in a family of host-guest pentapeptides. *Biochemistry* **35**, 5109–5124 (1996).
15. Jayasinghe, S., Hristova, K. & White, S. H. Energetics, stability, and prediction of transmembrane helices. *J. Mol. Biol.* **312**, 927–934 (2001).
16. Cornette, J. L. *et al.* Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685 (1987).
17. Degli Esposti, M., Crimi, M. & Venturoli, G. A critical evaluation of the hydropathy profile of membrane proteins. *Eur. J. Biochem.* **190**, 207–219 (1990).
18. Heinrich, S., Mothes, W., Brunner, J. & Rapoport, T. The SecYp complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102**, 233–244 (2000).
19. Lu, L. P. & Deber, C. M. Guidelines for membrane protein engineering derived from de novo designed model peptides. *Biopolymers* **47**, 41–62 (1998).
20. Bechinger, B. Membrane insertion and orientation of polyaniline peptides: A N-15 solid-state NMR spectroscopy investigation. *Biophys. J.* **81**, 2251–2256 (2001).

21. Lewis, R. N. *et al.* A polyaniline-based peptide cannot form a stable transmembrane alpha-helix in fully hydrated phospholipid bilayers. *Biochemistry* **40**, 12103–12111 (2001).
22. Wallin, E., Tsukihara, T., Yoshikawa, S., von Heijne, G. & Elofsson, A. Architecture of helix bundle membrane proteins: An analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci.* **6**, 808–815 (1997).
23. Killian, J. A. & von Heijne, G. How proteins adapt to a membrane-water interface. *Trends Biochem. Sci.* **25**, 429–434 (2000).
24. Yau, W. M., Wimley, W. C., Gawrisch, K. & White, S. H. The preference of tryptophan for membrane interfaces. *Biochemistry* **37**, 14713–14718 (1998).
25. Wimley, W. C. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* **3**, 842–848 (1996).
26. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142 (1984).
27. Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94**, 795–807 (1998).
28. McCormick, P. J., Miao, Y., Shao, Y., Lin, J. & Johnson, A. E. Cotranslational protein integration into the ER membrane is mediated by the binding of nascent chains to translocon proteins. *Mol. Cell* **12**, 329–341 (2003).
29. Presta, L. G. & Rose, G. D. Helix signals in proteins. *Science* **240**, 1632–1641 (1988).
30. Richardson, J. S. & Richardson, D. C. Amino acid preferences for specific locations at the ends of α -helices. *Science* **240**, 1648–1652 (1988).
31. Yohannan, S. *et al.* Proline substitutions are not easily accommodated in a membrane protein. *J. Mol. Biol.* **341**, 1–6 (2004).
32. Kuroiwa, T., Sakaguchi, M., Mihara, K. & Omura, T. Systematic analysis of stop-transfer sequence for microsomal membrane. *J. Biol. Chem.* **266**, 9251–9255 (1991).
33. Anthony, V. & Skach, W. R. Molecular mechanism of P-glycoprotein assembly into cellular membranes. *Curr. Protein Pept. Sci.* **3**, 485–501 (2002).
34. Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**, 187–208 (1999).
35. Liljeström, P. & Garoff, H. A new generation of animal cell expression vectors based on the Semliki Forest virus replicon. *Biotechnology* **9**, 1356–1361 (1991).
36. Liljeström, P., Lusa, S., Huylebroeck, D. & Garoff, H. *In vitro* mutagenesis of a full-length cDNA clone of Semliki Forest virus: the small 6,000-molecular-weight membrane protein modulates virus release. *J. Virol.* **65**, 4107–4113 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We wish to thank E. Missioux for technical assistance and R. MacKinnon, D. Rees, and T. Rapoport for comments. This work was supported by grants from the Swedish Cancer Foundation to G.v.H. and I.M.N., the Marianne and Marcus Wallenberg Foundation and the Swedish Research Council to G.v.H., the Magnus Bergvall Foundation to I.M.N., and the National Institute of General Medical Sciences to S.H.W.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.v.H. (gunnar@dbb.su.se).

Drosophila Spire is an actin nucleation factor

Margot E. Quinlan¹, John E. Heuser², Eugen Kerkhoff³ & R. Dyche Mullins¹

¹Department of Cellular and Molecular Pharmacology, UCSF Medical School, San Francisco, California 94107, USA

²Department of Cell Biology and Physiology, Washington University, St Louis, Missouri 63110, USA

³Institut für Medizinische Strahlenkunde und Zellforschung, Universität Würzburg, Würzburg 97078, Germany

The actin cytoskeleton is essential for many cellular functions including shape determination, intracellular transport and locomotion. Previous work has identified two factors—the Arp2/3 complex and the formin family of proteins—that nucleate new actin filaments via different mechanisms. Here we show that the *Drosophila* protein Spire represents a third class of actin nucleation factor. *In vitro*, Spire nucleates new filaments at a rate that is similar to that of the formin family of proteins but slower than in the activated Arp2/3 complex, and it remains associated with the slow-growing pointed end of the new filament. Spire contains a cluster of four WASP homology 2 (WH2) domains, each of which binds an actin monomer. Maximal nucleation activity requires all four WH2 domains along with an additional actin-binding motif, conserved among Spire proteins. Spire itself is conserved among metazoans and, together with the formin Cappuccino, is required for axis specification in oocytes and embryos, suggesting that multiple actin nucleation factors collaborate to construct essential cytoskeletal structures.

The *Drosophila* genes *spire* and *cappuccino* are required for proper development of oocytes and embryos. Mutations in either gene result in failure of the egg, and later the developing embryo, to establish dorsal–ventral and anterior–posterior axes of polarity¹. The molecular mechanism by which *spire* and *cappuccino* contribute to oocyte and embryo polarity is not understood but is thought to involve actin assembly. The phenotypes of *spire* and *cappuccino* mutations are mimicked by addition of cytochalasin D, a drug that inhibits actin polymerization, and by mutations in the actin-binding protein profilin². The *cappuccino* gene product, Cappuccino (Capu), contains formin homology domains (FH1 and FH2) similar to previously described FH domains that nucleate actin filament formation^{3,4}. The *spire* gene product, Spire (Spir), contains a cluster of four potential actin-monomer-binding sequences similar to WASP homology 2 (WH2) domains^{5,6} and to the amino-terminal portion of the β -thymosin domain⁷. Other proteins that contain WH2/ β -thymosin repeats have been shown to strongly inhibit spontaneous actin nucleation, probably by inhibiting interactions between monomers. In contrast, we find that Spir nucleates new filaments by a previously unknown mechanism.

Spir nucleates new actin filaments

Spir initially attracted our attention because it contains four putative WH2 domains and a stretch of acidic residues⁵—hallmarks of proteins that activate the Arp2/3 complex⁸. Spir does not, however, contain the central helical domain required for activation of the Arp2/3 complex by WASP-family proteins⁹, so we tested its effect on actin assembly *in vivo* and *in vitro*. Overexpression of Spir in NIH 3T3 cells induces formation of filamentous actin clusters⁵ (Fig. 1a, left panels; see also Supplementary Fig. S1) similar to those formed by overexpression of WASP-family proteins¹⁰ (Fig. 1a, right panels). Actin clusters induced by WASP-family proteins co-localize with the Arp2/3 complex, whereas Spir-induced clusters do not (Fig. 1a). Using truncation mutants we localized the actin-cluster-forming activity of Spir to the N-terminal half of the molecule; that is, the region containing the acidic domain and the four WH2 repeats (Supplementary Fig. S1). These data suggested that Spir might promote actin filament assembly via an Arp2/3-independent pathway.

To determine the effect of Spir on actin assembly *in vitro* we expressed the N-terminal half of the protein ((NT)Spir; amino acids

1–520) fused to a 6 $\times$ -His tag in *Escherichia coli* and purified it by Ni²⁺-affinity chromatography (see Supplementary Methods). We found that (NT)Spir alone induces rapid and dose-dependent actin polymerization in the absence of the Arp2/3 complex (Fig. 1b, c). Addition of the Arp2/3 complex does not significantly alter the rate of Spir-induced polymerization, suggesting that the two nucleators do not interact (Fig. 1b; see also Supplementary Information).

Actin filament assembly may be accelerated by one of three processes: (1) an increase in the rate of filament elongation; (2) generation of new filament ends via severing; or (3) *de novo* filament nucleation. At concentrations most effective for accelerating polymerization, (NT)Spir has no effect on elongation of pre-formed actin seeds (Supplementary Fig. S2). This result rules out the first and second processes (elongation and severing) and indicates that Spir nucleates actin filaments *de novo*. *In vivo* actin filaments assemble from their fast-growing, barbed ends¹¹. (NT)Spir-induced polymerization is inhibited by cytochalasin D, a factor that caps the fast-growing barbed end of actin filaments (Supplementary Fig. S3), indicating that Spir, like the Arp2/3 complex and the formins, generates filaments with elongating barbed ends. Similar to the activity of the Arp2/3 complex, Spir-dependent nucleation is slowed by the actin-monomer-binding protein profilin; however, even at a two-to-one ratio of profilin to actin, Spir's nucleation activity is significant (Supplementary Fig. S3b).

Spir differs from known nucleators

Spir has no sequence homology to the formin family of proteins or to subunits of the Arp2/3 complex. To understand better its nucleation mechanism we compared the activity of Spir to that of the activated Arp2/3 complex and the *Drosophila* formin Capu (Fig. 2a). We expressed and purified a Capu construct containing the FH1 and FH2 domains (Capu(FH1FH2)) and found that it nucleates actin filament formation (Supplementary Fig. S4). Under identical conditions, the maximal nucleation activity of (NT)Spir is approximately equal to that of Capu(FH1FH2) but significantly lower than that of the activated Arp2/3 complex (Fig. 2a). At low concentrations (<1 μ M) Spir is more active than Capu. At high concentrations of Spir (approaching 1:1 stoichiometry with actin) both the rate of polymerization and the steady-state concentration of actin polymer decrease (Figs 1c and 2a). Neither the Arp2/3 complex nor the formins exhibit such a biphasic dose response,

suggesting that the biochemical mechanism of Spir activity is significantly different (see also Supplementary Fig. S5).

The Arp2/3 complex nucleates new filaments from the sides of preformed filaments and generates arrays of branched actin filaments. We tested whether Spir or Capu interact with the sides of actin filaments. Neither Capu(FH1FH2) nor (NT)Spir co-pellet with filamentous actin (our unpublished observations), suggesting that they do not bind to the sides of actin filaments. In addition, micrographs of fluorescently labelled actin filaments nucleated by Capu(FH1FH2) or (NT)Spir indicate that neither protein generates dendritically branched actin filaments characteristic of the Arp2/3 complex (Fig. 2b)^{3,12,13}. Furthermore, addition of preformed filaments has no effect on the nucleation activity of (NT)Spir (Supplementary Fig. S2).

Similar to the Arp2/3 complex but unlike the formins, Spir binds the pointed end of actin filaments. Formins interact with the fast-growing barbed end of actin filaments and can alter the rate of filament elongation and disassembly from the barbed end^{3,14,15}, whereas the Arp2/3 complex caps the slow-growing pointed end with nanomolar affinity (Fig. 2c)¹³. (NT)Spir has no effect on elongation or shortening of actin filaments with free barbed ends (Supplementary Fig. S2); however, when the fast-growing barbed end is capped with gelsolin, Spir inhibits disassembly from the pointed end (Fig. 2c). Our data slightly underestimate the effect of Spir on pointed-end disassembly because, under our assay conditions, sequestration of monomeric actin by (NT)Spir (Supplementary Fig. S5a) should accelerate depolymerization. We

estimate that (NT)Spir binds pointed ends of actin filaments with an affinity in the micromolar range. Interestingly, actobindin and Ciboulot, which contain multiple WH2-like motifs, also weakly associate with pointed ends of actin filaments^{16,17}.

Spir WH2 domains assemble actin nuclei

Next we tested whether the WH2-like sequences in Spir bind monomeric actin and whether they participate in filament nucleation. We refer to these domains as WH2-A through to WH2-D (N-terminal to carboxy-terminal), and we refer to the intervening sequences as linker 1 through to linker 3 (L-1 to L-3, Fig. 3a). The sequences differ by varying degrees compared with canonical WASP-family WH2 sequences (Fig. 3a). WH2-B and WH2-D are the most highly conserved and were originally identified by ref. 6, which showed that WH2-D binds actin. We expressed and purified each Spir WH2 domain fused to glutathione S-transferase (GST) and tested its actin binding activity *in vitro*. Similar to conventional WH2 domains⁸, each Spir domain inhibits spontaneous polymerization (Supplementary Fig. S6a). Conventional WH2 domains contact actin via a set of highly conserved hydrophobic amino acid residues¹⁶ (A. E. Kelly and R.D.M., unpublished results), and mutation of these residues in the Spir WH2-like domains (Fig. 3a) abolishes their ability to inhibit polymerization (Supplementary Fig. S6b). Together, our results indicate that each WH2-like sequence in Spir interacts with actin in a manner similar to a traditional WH2 domain.

To determine the connection between actin binding by the WH2

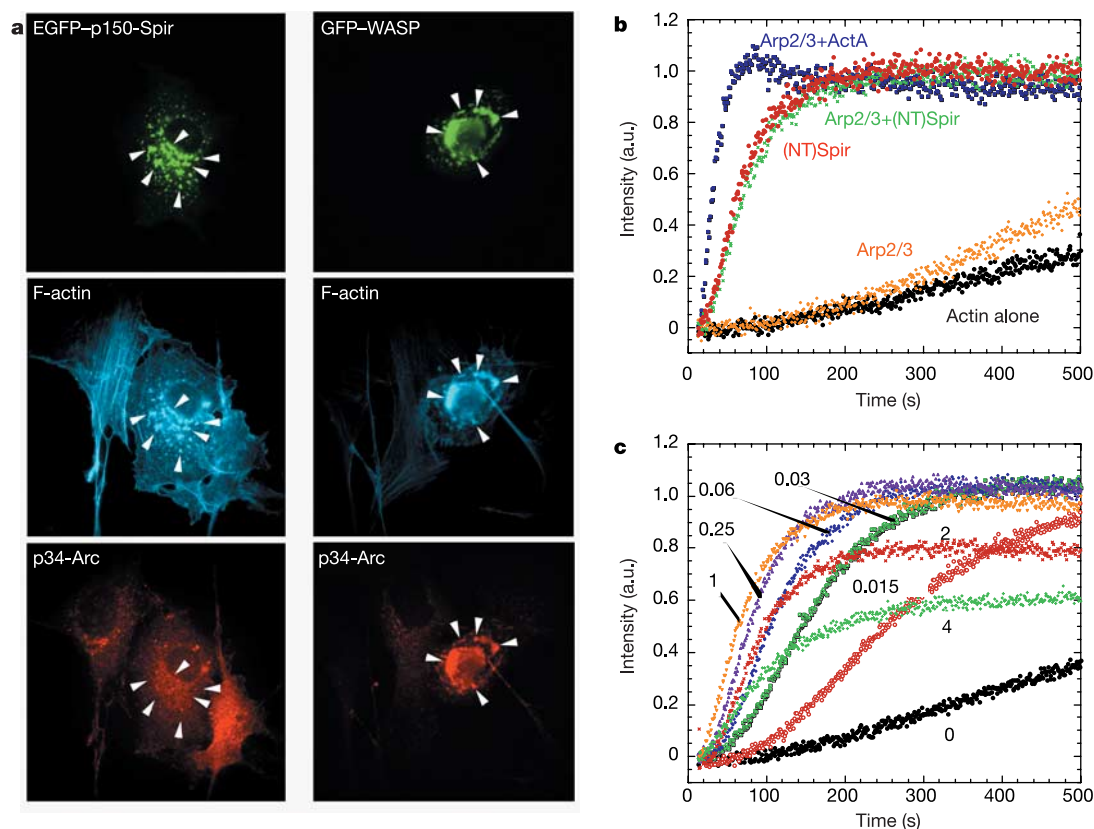


Figure 1 Spir nucleates actin filaments. **a**, Expression of either GFP-tagged Spir (EGFP-p150-Spir, green) or WASP (GFP-WASP, green) in NIH 3T3 cells leads to accumulation of filamentous actin (visualized with phalloidin, blue). The p34-Arc subunit (red) of the Arp2/3 complex co-localizes with WASP-induced actin clusters but not Spir-induced actin clusters. **b**, (NT)Spir nucleates actin independently of the Arp2/3 complex. Pyrene-actin polymerization assays with 0.5 μ M of (NT)Spir alone ((NT)Spir) are as fast as those with (NT)Spir and 20 nM Arp2/3 (Arp2/3 + (NT)Spir). Polymerization induced by 20 nM of

Arp2/3 alone (Arp2/3) and in the presence of saturating ActA³⁴ (Arp2/3 + ActA) are shown for comparison. These and all other polymerization traces are scaled (but not normalized) to an arbitrary plateau value of 1 for 0.5 μ M (NT)Spir. They were not normalized so as not to mask the effect of sequestration. **c**, Spir nucleates actin filament assembly in a dose-dependent manner. The indicated amount (in μ M) of (NT)Spir was added to pyrene-actin polymerization assays. See Supplementary Information for detailed Methods. a.u., arbitrary units.

domains and filament nucleation by (NT)Spir, we mutated the WH2 sequences individually and in combination (Fig. 3a, b). Mutation of the essential hydrophobic residues in individual WH2 domains produced nucleation defects of differing severity (Fig. 3b and Supplementary Table S1). Mutation of WH2-A produced the smallest effect ($t_{1/2} = 73 \pm 6$ s versus 67 ± 5 s for wild type in the presence of $4 \mu\text{M}$ actin; mean $\pm$ s.d.); mutation of WH2-B and WH2-C produced progressively greater effects ($t_{1/2} = 97 \pm 6$ s and 150 ± 10 s, respectively); and mutation of WH2-D produced the most profound effect of all on nucleation activity ($t_{1/2} = 330 \pm 20$ s). Removal of WH2 domains by truncation produced similar effects (Fig. 3c and Supplementary Table S1). A construct encompassing only the WH2 cluster (ABCD) displayed similar levels of nucleation activity as (NT)Spir (Fig. 3c). Removal of one or even two N-terminal WH2 domains had only a small effect whereas removal of WH2-D markedly reduced nucleation activity, and removal of both WH2-C and WH2-D (along with linkers 2 and 3; (NT)Spir[$\Delta 2\text{C}3\text{D}$]) nearly abolished nucleation activity (Fig. 3c and Supplementary Table S1). We detected only weak nucleation activity at high concentrations of (NT)Spir[$\Delta 2\text{C}3\text{D}$] (our unpublished observation).

A unique actin-binding site

Simultaneous mutation of all four WH2 domains ((NT)Spir [A*B*C*D*]) greatly diminishes nucleation activity but fails to abolish it ($t_{1/2} = 350 \pm 20$ s versus 420 ± 30 s for $4 \mu\text{M}$ actin alone; Fig. 3b). (NT)Spir, therefore, must contain at least one

more actin-binding site with weak nucleation activity. Because the nucleation activity of ABCD is identical to that of (NT)Spir, we focused our attention on the short (~ 15 amino acid) sequences between the WH2 domains (Supplementary Fig. S7). To determine their role in nucleation, we replaced each linker with a set of four flexible Gly-Ser repeats that maintain the same spacing between WH2 domains (see Supplementary Information). Replacing either L-1 or L-2 ((NT)Spir[gs1] and (NT)Spir[gs2]) produced negligible effects on nucleation activity (Fig. 3d). In contrast, replacement of L-3 ((NT)Spir[gs3]) markedly diminished the nucleation activity and replacement of all three linkers was equivalent to replacing L-3 alone (Fig. 3d). Thus L-3 appeared to be a likely candidate for an actin-binding domain involved in filament nucleation.

We expressed L-3 as a $6 \times$ His-tagged peptide (Fig. 3a and Supplementary Information) and tested its activity *in vitro*. By itself, the L-3 peptide weakly nucleates actin assembly (Fig. 3e) with activity similar to that of (NT)Spir[A*B*C*D*]. Because of its size we assume that L-3 promotes actin dimer formation by stabilizing the interaction of two actin monomers. The sequence seems to be unrelated to other known actin-binding motifs and database searches yield only Spir-like proteins (Fig. 3a; see also Supplementary S7).

Ultrastructure of the Spir-actin complex

We hypothesize that Spir forms a new actin filament by first binding several actin monomers and then assembling them stereospecifically into a filament nucleus: a dimer, trimer, or a tetramer. To test this

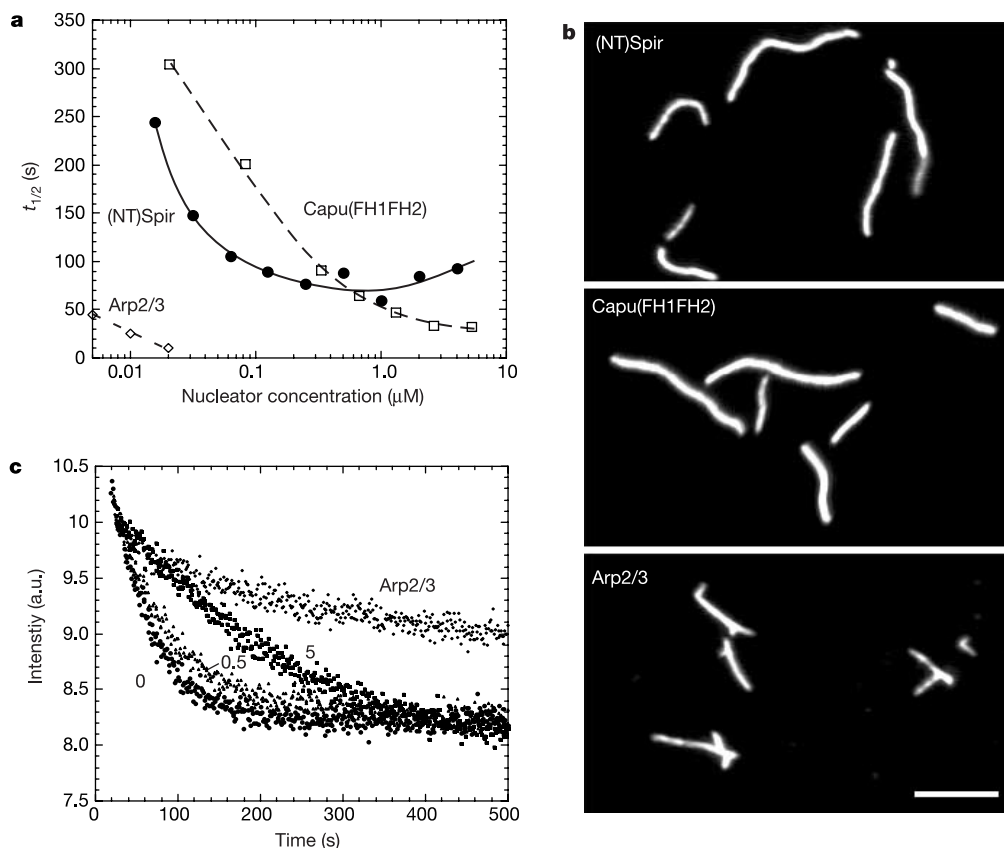


Figure 2 Spir, Capu and the Arp2/3 complex nucleate polymerization by distinct mechanisms. **a**, To compare activity levels of different nucleators, we plotted the time to half-maximal polymerization ($t_{1/2}$) as a function of (NT)Spir (filled circles), Capu(FH1FH2) (open squares) and Arp2/3 complex (open diamonds) concentrations. (NT)Spir and Capu(FH1FH2) nucleate at similar rates whereas filament formation by the Arp2/3 complex is significantly faster. **b**, Filaments produced by $0.5 \mu\text{M}$ (NT)Spir (top panel) or

$0.5 \mu\text{M}$ Capu(FH1FH2) (middle panel) are not crosslinked, whereas those produced by the Arp2/3 complex (20 nM plus ActA) are branched in a characteristic dendritic pattern (bottom panel). **c**, Depolymerization of gelsolin-capped filaments in the presence of 0, 0.5 and $5 \mu\text{M}$ (NT)Spir or 100 nM Arp2/3 shows that (NT)Spir caps the pointed end of filaments.

hypothesis, we performed electron microscopy on rotary-shadowed samples of (NT)Spir bound to actin. We combined (NT)Spir and actin at a 1:8 molar ratio under non-polymerizing conditions. We then diluted the samples into polymerization buffer and rapidly fixed them. Our protocol was designed to capture the maximum number of transient nucleating structures before they became bona fide actin filaments. In many fields we did, in fact, see actin filaments (Fig. 4b, lower left panels) but in almost all fields we also observed a novel, rod-shaped structure 22 ± 4 nm in length and 6.5 ± 0.3 nm in diameter ($n = 43$; Fig. 4). We observed Spir-actin rods of the

same size ($6.5 \pm 1 \times 24 \pm 3$ nm, $n = 17$) in the presence of latrunculin B (Fig. 4b, lower right), which blocks actin filament formation but does not prevent binding of actin monomers to WH2 domains^{16,18}; however, we never observed such structures in samples of actin or (NT)Spir alone. These Spir-actin rods were distinguishable from short actin filaments by their uniform length and smaller diameter (6.5 nm compared with 7.5 nm). The length and diameter of Spir-actin rods is consistent with four actin monomers aligned along their long axes. When we mixed the WH2 cluster (ABCD) with actin it produced rod-shaped structures nearly identical to

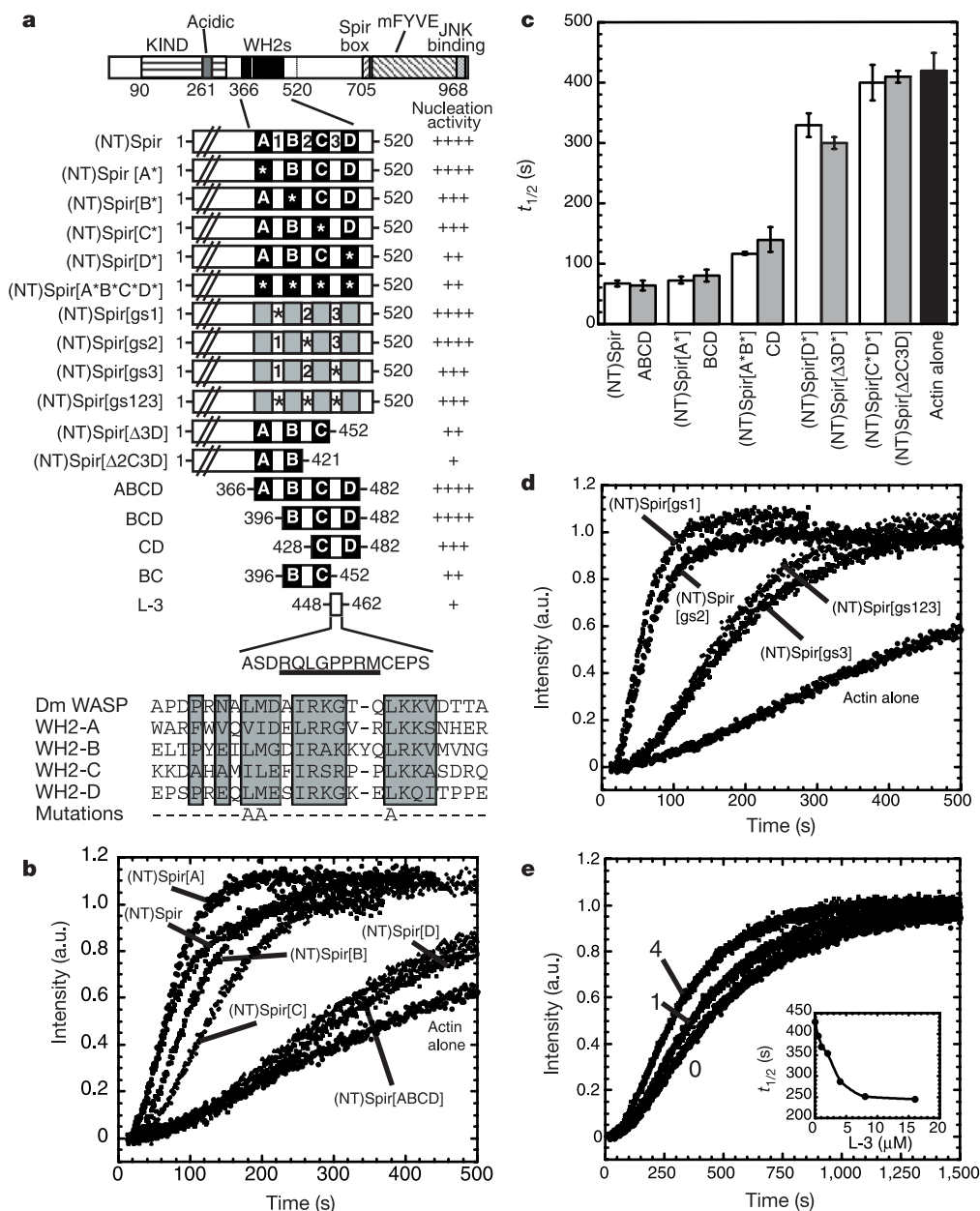


Figure 3 Molecular dissection of Spir. **a**, Diagram of full-length Spir and mutants generated in this study. The domains identified are described in Supplementary Fig. S7. The sequence of L-3 is shown with the eight residues replaced by four Gly/Ser repeats underlined. The WH2 domains are aligned with the *Drosophila* WASP WH2 domain, and the alanine mutations are shown at the bottom. Nucleation activity of each Spir mutant (0.5 μ M with 4 μ M actin) is indicated in the right column. Polymerization stimulated by wild-type (NT)Spir is denoted by +++; nucleation equivalent to spontaneous actin polymerization is indicated by +. **b**, Each WH2 domain has a different role in nucleation; for example, mutation of WH2-A ((NT)Spir[A*]) had almost no effect on nucleation,

whereas mutation of WH2-D ((NT)Spir[D*]) markedly decreased activity. **c**, Comparison of $t_{1/2}$ induced by truncation mutants. Activity of (NT)Spir and actin alone (black column) are shown at either end of the graph. Corresponding pairs of truncation (grey) and point mutants (white) are plotted side by side. Error bars are $\pm$ s.d. **d**, The sequences linking WH2 domains have different roles in actin nucleation. Only replacing L-3 ((NT)Spir[gs3]) markedly reduced nucleation activity. **e**, L-3 alone weakly nucleates filament formation. Concentrations of L-3 (in μ M) are indicated. Inset: $t_{1/2}$ versus L-3 concentration. Note that this domain does not sequester monomeric actin at high concentrations.

those produced by (NT)Spir ($5.9 \pm 0.9 \times 26 \pm 5$ nm, $n = 111$), whereas CD and actin produced structures that were approximately half as long ($5 \pm 1 \times 9 \pm 3$ nm, $n = 85$; Supplementary Fig. S8). These data are consistent with a model in which the WH2 domains bind actin monomers and align them end-to-end, as if along one strand of the long-pitch, two-start filament helix.

Mechanism of filament nucleation by Spir

Our results indicate that the C-terminal half of the Spir WH2 cluster—composed of WH2-C, L-3 and WH2-D—is the functional core of the protein. The main kinetic barrier to nucleation is formation of an actin dimer. We propose that Spir assembles an actin dimer when WH2-C and WH2-D each bind an actin monomer and L-3 coordinates their interaction (Supplementary Fig. S9). Similar to the first dimer formed during spontaneous nucleation¹⁹, we expect this dimer to lie along one strand of the long-pitch actin helix. On the basis of their effects on nucleation and our own electron microscopy data, we propose that WH2-B and WH2-A add a third and fourth monomer to the initial dimer (Fig. 5b).

In addition to electron microscopy (Fig. 4), spatial constraints

imposed by the Spir sequence¹⁶ (Fig. 5a) and the atomic structure of the WH2-like domain of Ciboulot¹⁶ suggest that Spir stacks monomers along one strand of the long-pitch filament helix (Fig. 5b). The N-terminal portion of a WH2 domain would block addition of the next monomer at the barbed end. We propose that, similar to Ciboulot and the WASP-family WH2 domains, the N-terminal portion of the Spir WH2 domains dissociates rapidly upon incorporation into the nascent nucleus. Consistent with this idea, residues in the Ciboulot actin-binding site important for dissociation and promotion of actin filament assembly are conserved in Spir (Supplementary Fig. S7c). The only structure consistent with the lengths of the linker sequences is a linear arrangement of the four WH2-bound monomers, with WH2-D at the pointed end and WH2-A at the barbed end. Binding of an additional monomer to the interface between any of the WH2-bound monomers (asterisks in Fig. 5b) would result in formation of a stable nucleus and rapid filament elongation.

The fact that Spir assembles nuclei even when all three linker sequences are mutated ((NT)Spir[gs123]) suggests that tethering multiple, WH2-bound actin monomers in close proximity may be

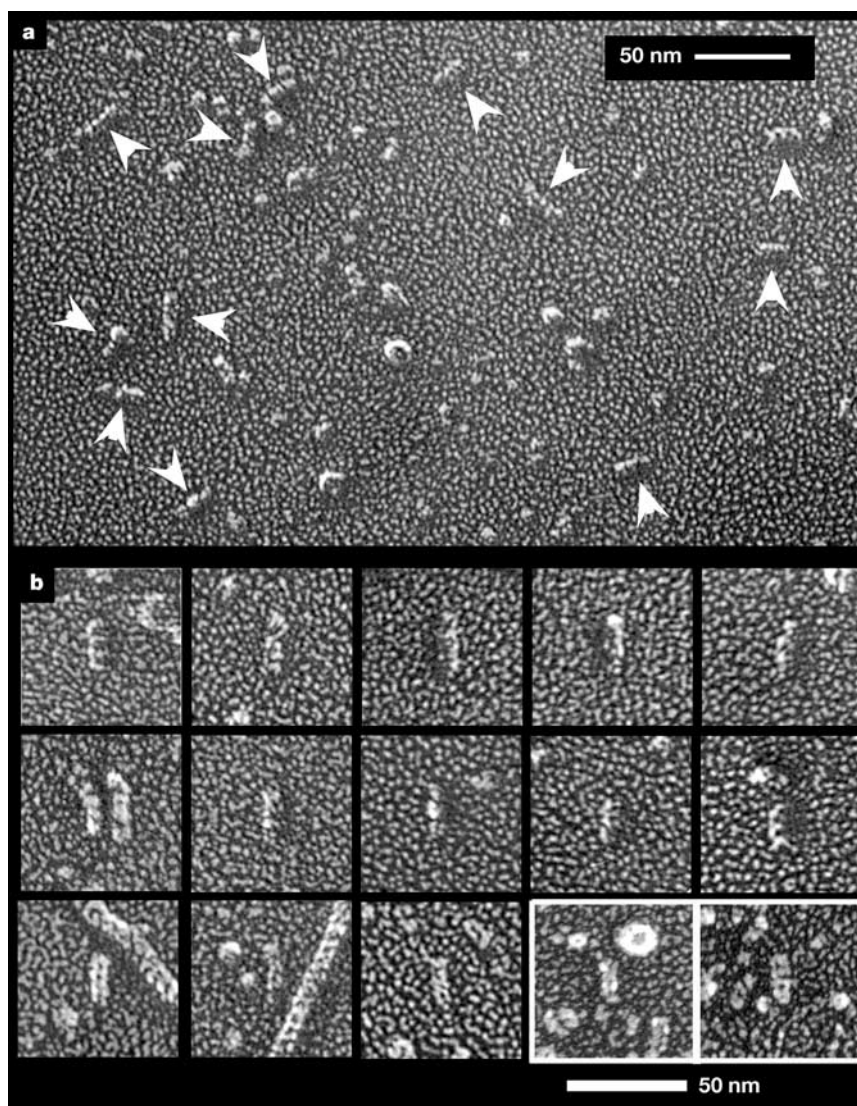


Figure 4 Spir–actin complexes are rod-like structures. **a**, Sample electron micrograph of actin and (NT)Spir (1:8 ratio) rapidly fixed after polymerizing conditions are introduced. The field contains multiple examples of an $\sim 22 \times 6.5$ nm rod-shaped structure observed only in samples containing both (NT)Spir and actin. **b**, Individual Spir–actin rod complexes

observed under both polymerizing and non-polymerizing conditions. The two examples in the white box at the lower right show similar structures seen in the presence of the polymerization inhibitor, latrunculin B. Typical actin filaments next to rods are seen in the two lower-left boxes.

sufficient to promote filament formation. The only other known tandem repeats of competent WH2 domains are found in N-WASP and tetra-thymosin- β . By themselves, N-WASP and tetra-thymosin- β allow elongation of the barbed ends of filaments, but strongly inhibit spontaneous nucleation. Further work is required to determine whether WH2 domains in Spir are specially adapted to promote nucleation, or whether sequences in other WH2-containing proteins are specially adapted to prevent nucleation, as in the case of thymosin- β 4 (ref. 20).

Cellular function of Spir

After the Arp2/3 complex and the formins, Spir is the third actin nucleation factor to be discovered. Why do cells require more than one mechanism for constructing actin filaments? The answer probably lies in the diversity of functions performed by the actin cytoskeleton. Different functions require actin networks with different architectures, and the architecture of an actin network is determined in part by the mechanism of filament nucleation. Dendritic nucleation by the Arp2/3 complex, for example, produces space-filling actin networks capable of resisting mechanical deformation. This activity is required for amoeboid motility, phagocytosis and intracellular motility of endosomal vesicles and some pathogens^{21,22}. The formins do not crosslink new filaments into branched arrays but remain attached to their growing ends and probably tether them to specific locations^{15,23}. Unbranched fila-

ments generated by formins are essential for construction of actin cables in budding yeast, and stress fibres and contractile rings in mammalian cells^{24–26}.

The regulation and expression pattern of Spir differ from those of the other nucleators. Unlike the Arp2/3 complex and the formins, Spir has no obvious orthologues in any sequenced protozoan genome; however, it is highly conserved across metazoan species (Supplementary Fig. S5a). Two mammalian isoforms, Spir-1 and Spir-2, are widely expressed in embryonic tissues but limited primarily to the central nervous system of adults²⁷. The Arp2/3 complex and the Diaphanous-related formins are downstream of Rho-family G proteins, whereas Spir and the mammalian homologue of Capu, formin-1, are MAP kinase substrates^{5,28}. Spir proteins are targeted to intracellular membranes by a C-terminal-modified FYVE zinc finger motif, and co-localize with the GTPase Rab11, which is involved in vesicle transport processes^{29,30}. As with *spir* and *capu*, *Drosophila Rab11* belongs to the posterior group of genes^{31,32}. In addition, *Rab11*, *spir* and *capu* mutants have similar defects in microtubule plus-end orientation during oogenesis^{6,32,33}. These data together with our results suggest that Spir has evolved specifically to construct actin-based structures required for polarization of multicellular organisms. □

Methods

We cloned the *spir* gene and generated mutants using standard techniques (see Supplementary Methods). We expressed wild-type and mutant Spir proteins in *E. coli* and purified them using a combination of Ni²⁺-affinity and anion exchange chromatography (details in Supplementary Methods). Immuno-labelling and fluorescence microscopy were performed using standard methods (see Supplementary Methods).

Actin kinetics

All pyrene-actin polymerization assays were performed using 4 μ M monomeric actin (5% pyrene labelled) in KMEI (50 mM KCl, 1 mM MgCl₂, 1 mM EGTA, 10 mM imidazole pH 7.0). Ca-actin was converted to Mg-actin by incubation in ME (50 mM MgCl₂, 0.2 mM EGTA) for 2 min before addition of KMEI and test components. Unless otherwise indicated, 0.5 μ M nucleator ((NT)Spir or mutant) was added. Fluorescence was excited at 365 nm and detected at 407 nm using an ISS PC1/K2 fluorimeter.

Depolymerization assays were performed by diluting 10 μ M filamentous actin (10% pyrene labelled) to 1 μ M in KMEI. Gelsolin-capped filaments were created by polymerizing actin off of gelsolin-actin seeds (ref. 13).

Electron microscopy

We first combined actin and (NT)Spir at an 8:1 molar ratio in buffer A. To capture early stages of nucleation, we rapidly diluted samples to 5 μ g ml⁻¹ in KMEI, adsorbed them to mica flakes at 25 °C, then quick-froze and freeze-dried them. Details regarding image acquisition and calibration are given in Supplementary Methods.

Received 17 September; accepted 29 November 2004; doi:10.1038/nature03241.

- Manseau, L. J. & Schupbach, T. cappuccino and spire: two unique maternal-effect loci required for both the anteroposterior and dorsoventral patterns of the *Drosophila* embryo. *Genes Dev.* **3**, 1437–1452 (1989).
- Manseau, L., Calley, J. & Phan, H. Profilin is required for posterior patterning of the *Drosophila* oocyte. *Development* **122**, 2109–2116 (1996).
- Pruyne, D. *et al.* Role of formins in actin assembly: nucleation and barbed-end association. *Science* **297**, 612–615 (2002).
- Evangelista, M., Pruyne, D., Amberg, D. C., Boone, C. & Bretscher, A. Formins direct Arp2/3-independent actin filament assembly to polarize cell growth in yeast. *Nature Cell Biol.* **4**, 260–269 (2002).
- Otto, I. M. *et al.* The p150-Spir protein provides a link between c-Jun N-terminal kinase function and actin reorganization. *Curr. Biol.* **10**, 345–348 (2000).
- Wellington, A. *et al.* Spire contains actin binding domains and is related to ascidian posterior end mark-5. *Development* **126**, 5267–5274 (1999).
- Paunola, E., Mattila, P. K. & Lappalainen, P. WH2 domain: a small, versatile adapter for actin monomers. *FEBS Lett.* **513**, 92–97 (2002).
- Higgs, H. N., Blanchoin, L. & Pollard, T. D. Influence of the C terminus of Wiskott-Aldrich syndrome protein (WASP) and the Arp2/3 complex on actin polymerization. *Biochemistry* **38**, 15212–15222 (1999).
- Panchal, S. C., Kaiser, D. A., Torres, E., Pollard, T. D. & Rosen, M. K. A conserved amphipathic helix in WASP/Scar proteins is essential for activation of Arp2/3 complex. *Nature Struct. Biol.* **10**, 591–598 (2003).
- Abo, A. Understanding the molecular basis of Wiskott-Aldrich syndrome. *Cell. Mol. Life Sci.* **54**, 1145–1153 (1998).
- Devreotes, P. N. & Zigmond, S. H. Chemotaxis in eukaryotic cells: a focus on leukocytes and *Dictyostelium*. *Annu. Rev. Cell Biol.* **4**, 649–686 (1988).
- Blanchoin, L. *et al.* Direct observation of dendritic actin filament networks nucleated by Arp2/3 complex and WASP/Scar proteins. *Nature* **404**, 1007–1011 (2000).
- Mullins, R. D., Heuser, J. A. & Pollard, T. D. The interaction of Arp2/3 complex with actin: nucleation,

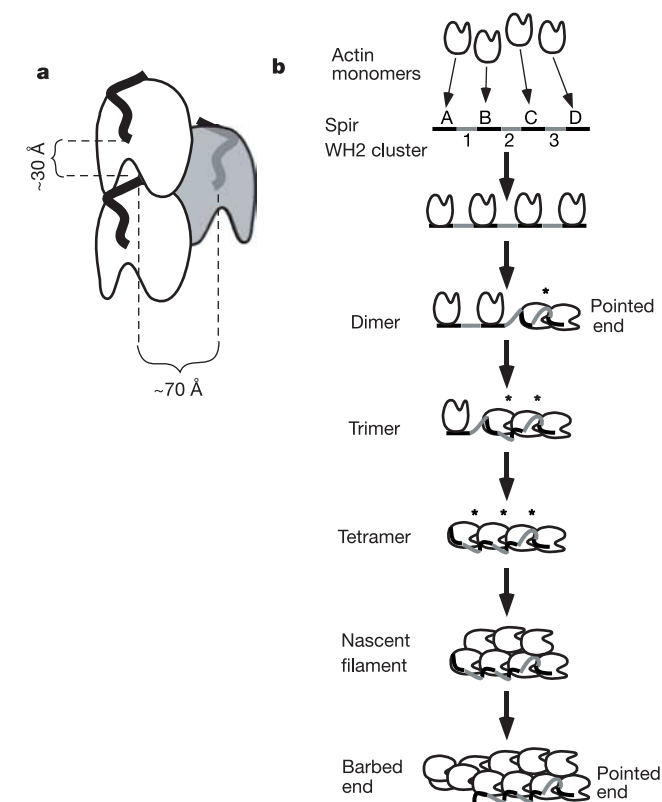


Figure 5 Mechanism of nucleation by Spir. We propose that Spir coordinates association of actin monomers along one strand of the two-start filament helix. **a**, Schematic alignment of WH2 domains on an actin trimer based on ref. 16. A 15 amino acid linker cannot connect the cross-filament dimer in this orientation. **b**, Up to four actin monomers can bind the WH2 cluster of Spir. Once monomers are bound to WH2-C and WH2-D they are coordinated and stabilized in a longitudinal dimer by L-3. Monomers from WH2-B and WH2-A are further stacked on this 'pre-nucleus' structure. Asterisks indicate positions where addition of a monomer would result in formation of a complete nucleus and subsequent rapid polymerization.

- high affinity pointed end capping, and formation of branching networks of filaments. *Proc. Natl Acad. Sci. USA* **95**, 6181–6186 (1998).
14. Kovar, D. R., Kuhn, J. R., Tichy, A. L. & Pollard, T. D. The fission yeast cytokinesis formin Cdc12p is a barbed end actin filament capping protein gated by profilin. *J. Cell Biol.* **161**, 875–887 (2003).
 15. Zigmond, S. H. Formin-induced nucleation of actin filaments. *Curr. Opin. Cell Biol.* **16**, 99–105 (2004).
 16. Hertzog, M. *et al.* The β -thymosin/WH2 domain; structural basis for the switch from inhibition to promotion of actin assembly. *Cell* **117**, 611–623 (2004).
 17. Hertzog, M., Yarmola, E. G., Didry, D., Bubbs, M. R. & Carlier, M. F. Control of actin dynamics by proteins made of β -thymosin repeats: the actobindin family. *J. Biol. Chem.* **277**, 14786–14792 (2002).
 18. Dayel, M. J. & Mullins, R. D. Activation of Arp2/3 complex: Addition of the first subunit of the new filament by a WASP protein triggers rapid ATP hydrolysis on Arp2. *PLoS Biol.* **2**, E91 (2004).
 19. Sept, D. & McCammon, J. A. Thermodynamics and kinetics of actin filament nucleation. *Biophys. J.* **81**, 667–674 (2001).
 20. Irobi, E. *et al.* Structural basis of actin sequestration by thymosin- β 4: implications for WH2 proteins. *EMBO J.* **23**, 3599–3608 (2004).
 21. Pollard, T. D. & Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* **112**, 453–465 (2003).
 22. Fehrenbacher, K., Huckaba, T., Yang, H. C., Boldogh, I. & Pon, L. Actin comet tails, endosomes and endosymbionts. *J. Exp. Biol.* **206**, 1977–1984 (2003).
 23. Kovar, D. R. & Pollard, T. D. Insertional assembly of actin filament barbed ends in association with formins produces piconewton forces. *Proc. Natl Acad. Sci. USA* **101**, 14725–14730 (2004).
 24. Watanabe, N., Kato, T., Fujita, A., Ishizaki, T. & Narumiya, S. Cooperation between mDia1 and ROCK in Rho-induced actin reorganization. *Nature Cell Biol.* **1**, 136–143 (1999).
 25. Sagot, I., Klee, S. K. & Pellman, D. Yeast formins regulate cell polarity by controlling the assembly of actin cables. *Nature Cell Biol.* **4**, 42–50 (2002).
 26. Pelham, R. J. & Chang, F. Actin dynamics in the contractile ring during cytokinesis in fission yeast. *Nature* **419**, 82–86 (2002).
 27. Schumacher, N., Borawski, J. M., Leberfinger, C. B., Gessler, M. & Kerkhoff, E. Overlapping expression pattern of the actin organizers Spir-1 and formin-2 in the developing mouse nervous system and the adult brain. *Gene Expr. Patterns* **4**, 249–255 (2004).
 28. O'Rourke, D. A. *et al.* Hepatocyte growth factor induces MAPK-dependent formin IV translocation in renal epithelial cells. *J. Am. Soc. Nephrol.* **11**, 2212–2221 (2000).
 29. Kerkhoff, E. *et al.* The Spir actin organizers are involved in vesicle transport processes. *Curr. Biol.* **11**, 1963–1968 (2001).
 30. Sonnichsen, B., De Renzis, S., Nielsen, E., Rietdorf, J. & Zerial, M. Distinct membrane domains on endosomes in the recycling pathway visualized by multicolor imaging of Rab4, Rab5, and Rab11. *J. Cell Biol.* **149**, 901–914 (2000).
 31. Jankovics, F., Sinka, R. & Erdelyi, M. An interaction type of genetic screen reveals a role of the Rab11 gene in oskar mRNA localization in the developing *Drosophila melanogaster* oocyte. *Genetics* **158**, 1177–1188 (2001).
 32. Dollar, G., Struckhoff, E., Michaud, J. & Cohen, R. S. Rab11 polarization of the *Drosophila* oocyte: a novel link between membrane trafficking, microtubule organization, and oskar mRNA localization and translation. *Development* **129**, 517–526 (2002).
 33. Emmons, S. *et al.* Cappuccino, a *Drosophila* maternal effect gene required for polarity of the egg and embryo, is related to the vertebrate limb deformity locus. *Genes Dev.* **9**, 2482–2494 (1995).
 34. Welch, M. D., Rosenblatt, J., Skoble, J., Portnoy, D. A. & Mitchison, T. J. Interaction of human Arp2/3 complex and the *Listeria monocytogenes* ActA protein in actin filament nucleation. *Science* **281**, 105–108 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by grants from the NIH, Pew Charitable Trust and the Sandler Family Supporting Foundation (to R.D.M.). E.K. was supported by the Deutsche Forschungsgemeinschaft and the Wilhelm Sander-Stiftung. We thank C. B. Leberfinger and J. M. Borawski for technical assistance, and members of the Mullins laboratory for technical help and for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.D.M. (dyche@mullinslab.ucsf.edu).

Earth-mass dark-matter haloes as the first structures in the early Universe

J. Diemand*, B. Moore & J. Stadel

Institute for Theoretical Physics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

* Present address: Department of Astronomy and Astrophysics, University of California, 1156 High Street, Santa Cruz, California 95064, USA

The Universe was nearly smooth and homogeneous before a redshift of $z = 100$, about 20 million years after the Big Bang¹. After this epoch, the tiny fluctuations imprinted upon the matter distribution during the initial expansion began to collapse because of gravity. The properties of these fluctuations depend on the unknown nature of dark matter^{2–4}, the determination of which is one of the biggest challenges in present-day science^{5–7}. Here we report supercomputer simulations of the concordance cosmological model, which assumes neutralino dark matter (at present the preferred candidate), and find that the first objects to form are numerous Earth-mass dark-matter haloes about as large as the Solar System. They are stable against gravitational disruption, even within the central regions of the Milky Way. We expect over 10^{15} to survive within the Galactic halo, with one passing through the Solar System every few thousand years. The nearest structures should be among the brightest sources of γ -rays (from particle–particle annihilation).

The cosmological parameters of our Universe and initial conditions for structure formation have recently been measured via a combination of observations, including the cosmic microwave background (CMB)⁸, distant supernovae^{9,10} and the large-scale distribution of galaxies¹¹. Cosmologists now face the outstanding problem of understanding the origin of structure in the Universe from its strange mix of particles and vacuum energy^{12,13}.

Most of the mass of the Universe must be made up of a kind of non-baryonic particle^{1,14} that remains undetected in laboratory experiments. The leading candidate for this ‘dark matter’ is the neutralino, the lightest supersymmetric particle, which is predicted to solve several key problems in the standard model for particle physics⁵. This cold dark matter (CDM) candidate is not completely collisionless. It can collide with baryons, thus revealing its presence in laboratory detectors, although the cross-section for this interaction is extremely small. In a cubic-metre detector containing $\sim 10^{30}$ baryon particles, only a few collisions per day are expected from the $\sim 10^{13}$ dark-matter particles that flow through the experiment as the Earth moves through the Galaxy. The neutralino is its own anti-particle, and can self-annihilate, creating a shower of new particles including γ -rays⁵. The annihilation rate increases as the density squared; the central regions of the Galaxy and its satellites will therefore give the strongest signal^{15–18}. However, the expected rate is very low—the flux of photons on Earth is the same as we would receive from a single candle placed on Pluto. Numerous experiments using these effects are under way that may detect the neutralino within the next decade⁷. Furthermore, in the next few years the Large Hadron Collider (LHC) at CERN will confirm or rule out the concepts of supersymmetry (SUSY)⁶.

We followed the growth and subsequent gravitational collapse and virialization of the first structures in the CDM Universe with supercomputer calculations. The challenge is accurately to follow the evolution of the Universe on scales that are many orders of magnitude smaller than previously studied, while also capturing the gravitational dynamics from large scales. We use a multiscale technique¹⁹ to achieve the desired resolution within a small average-

density patch of the Universe that is nested within a hierarchy of larger and lower resolution grids of particles.

The fluctuations are imposed on the particles using accurate calculations of the linear theory power spectrum for a SUSY model with a particle mass $m_\nu = 100$ GeV. This includes collisional damping, free streaming and the transfer of fluctuations through the matter-radiation era of the Universe^{2–4}. The resulting power spectrum is close to a power law of $P(k) \propto k^n$ with $n = -3$, with an exponential cut-off at 0.6 co-moving parsecs, which corresponds to a mass scale of $10^{-6}M_\odot$, where $M_\odot$ is the mass of the Sun. The cut-off scale depends on the neutralino mass and decoupling energy. From accelerator searches we know that $m_\nu > 37$ GeV and that the cosmic matter density sets an upper limit at 500 GeV. The damping scale for the allowed neutralino models differ from the model we used by less than a factor of three in mass^{2–4} and structure formation is therefore very similar in all SUSY-CDM scenarios. A less popular CDM candidate is the axion, which has a much smaller damping scale of $10^{-13}M_\odot$. For comparison we simulated the high-resolution region with an axion CDM fluctuation spectrum on the resolved scales. Both models produce equal halo abundances above $5 \times 10^{-6}M_\odot$, but the axion model also forms bound structures down to the smallest resolved scales; see Fig. 3. Here we concentrate on the properties of the first structures to form in the SUSY-CDM model.

We evolve the initial particle distribution using a parallel multi-stepping tree code, starting at a redshift of $z = 350$ when the fluctuations are still linear. The high-resolution region forms the first nonlinear structures at $z = 60$ and the entire region quickly becomes distorted by the complex tidal field from the surrounding overdensities. By a redshift of $z = 26$, the high-resolution region begins to merge into the lower-resolution surroundings and we do not analyse the region further—however, this is sufficiently late that about 5% of the mass in the region has collapsed into bound dense structures (haloes); see Fig. 1.

The first dark-matter haloes to collapse and virialize are smooth triaxial objects of mass $10^{-6}M_\odot$ and half-mass radii of 10^{-2} pc. Figure 2 shows the density profiles of three representative haloes at $z = 26$ that are well fitted by single power-law density profiles $\rho(r) \propto r^{-\gamma}$ with slopes γ in the range from 1.5 to 2, similar to galactic haloes shortly after their formation²⁰. We note that the densities at the virial radius are about an order of magnitude higher than the density at $0.01r_{\text{virial}}$ in a galactic halo today, which makes the survival of many of these haloes as galactic substructure possible. The central resolved densities reach 10^9 times the mean background density at 1% of their virial radii. Unlike galactic and cluster-mass CDM haloes, they do not contain substructure because no smaller-mass haloes have collapsed in the hierarchy.

Figure 3 shows the mass function of haloes. We use a ‘friends of friends’ algorithm with a linking length set to identify the dense central regions of collapsed haloes, then for each halo centre we recursively search for the radius r_{200} that is at an overdensity of 200 times the cosmic mean density. The resulting halo mass function is steep: $dn(M)/d\log M \propto M^{-1}$. For comparison we plot an extrapolation of the halo mass function found on much larger scales $> 10^7M_\odot$ (ref. 21), which fits surprisingly well up to the cut-off scale of $10^{-6}M_\odot$, below which we find no more structures.

At these epochs the baryons are kept sufficiently warm by the CMB that they are unable to cool and form visible objects such as stars or planets within such tiny systems¹³. The dark haloes may be detected via gravitational effects such as lensing or dynamical perturbations. Although we cannot simulate the entire galactic halo at the resolution required to determine the survival statistics of these objects, we can make some simple estimates of their survival and abundances. As the Galactic halo is assembled, these first objects merge successively into more massive systems. From scales of $10^7M_\odot$ to $10^{15}M_\odot$ the mass function of substructure is a self-similar power law of slope $dn(m)/dm \propto m^{-1.9}$ (ref. 22). Extra-

polating the subhalo mass function to the smallest scales gives us a total number of substructure haloes $N(M > 10^{-6}M_{\odot}) \approx 5 \times 10^{15}$ and the expected number density of subhaloes at the solar radius is $n(R_{\odot}) \approx 500 \text{ pc}^{-3}$, assuming that they trace the mass. Although this extrapolation is made to much smaller masses than simulated previously, the substructure within haloes collapsing at $z \approx 15$ with masses $\sim 10^7 M_{\odot}$ fits the extrapolation from larger mass scales²³, even though they form from regions of the CDM power spectra with effective index $n \approx -2.95$.

Can these structures survive the strong disruptive gravitational forces from the Galaxy? The tidal radius is simply the inner Lagrange

point of the rotation of the two-body system. For haloes with power-law density profiles $\rho(r) \propto r^{-2}$ we find $r_t = (Rv_{\text{sat}})/(\sqrt{2}V_{\text{parent}})$ where v_{sat} and V_{parent} are the effective circular velocities ($V = \sqrt{GM/r}$) of the satellite and main halo, R is the pericentric distance of the satellite and G is the gravitational constant. For the smallest mini-haloes $v_{\text{sat}} \approx 1 \text{ m s}^{-1}$ and $r_{200} = 0.01 \text{ pc}$. Therefore, within the Galactic potential these haloes could survive completely intact to about 3 kpc from the centre, well within the galacto-centric position of the Sun. Encounters between haloes and with stars and molecular clouds may disrupt a small fraction of these structures, but using the impulsive heating approximation, we estimate that most will survive with little mass loss.

A significant fraction of the mass may lie within bound structures at our location within the Galaxy, lowering the available smoothly distributed matter necessary for direct detection experiments. The Earth passes through a dark-matter mini-halo every 10,000 years, an encounter which lasts for about 50 years, so that most of the time the Earth is within an underdense region of dark matter. Integrating the mass function from $10^{-6}M_{\odot}$ to $10^{10}M_{\odot}$, normalized such that 10% of the mass is within the substructure above a mass scale of $10^7 M_{\odot}$ (as given by simulations of Galactic haloes), we find that about 50% of the mass is bound to dark-matter substructures. The velocity perturbation to a planetary orbit or satellite is very small ($\sim 10^{-6} \text{ m s}^{-1}$), well below the observational constraints. However, resonant encounters and the cumulative effects of about 10^6 impulsive encounters may cause significant perturbations to some of the bodies orbiting in the Oort cloud surrounding the Solar System.

Compact objects in the mass range considered here could produce a microlensing signal in a multiply lensed quasar image, such as time-varying flux differences²⁴. The lensing object needs to be smaller than the Einstein radius (in centimetres):

$$r_E = 3.7 \times 10^{16} \sqrt{\frac{M}{hM_{\odot}}} \quad (1)$$

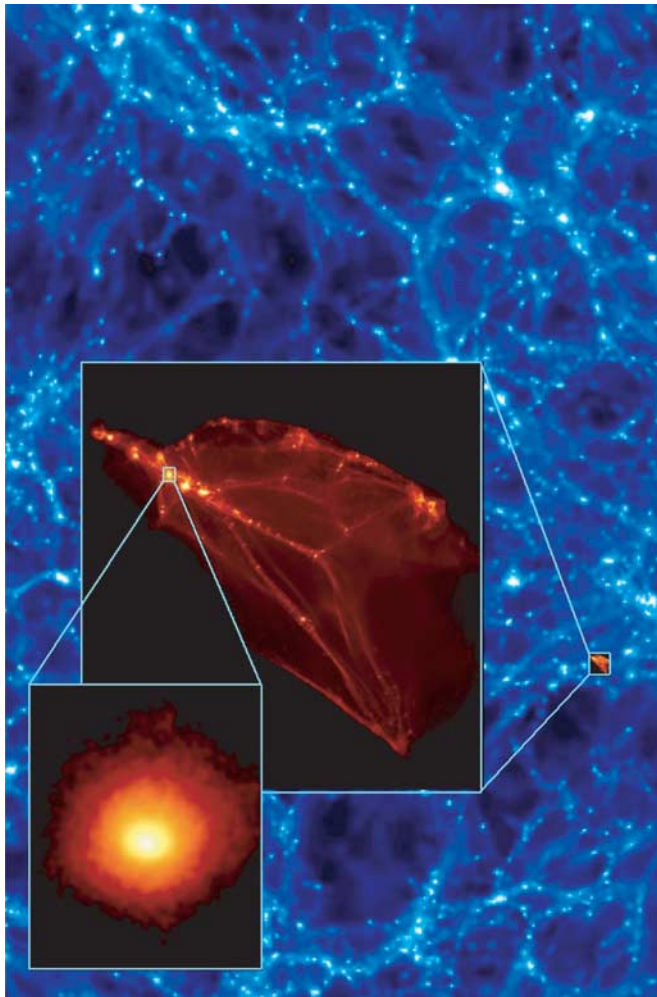


Figure 1 A zoom into one of the first objects to form in the Universe. The colours show the density of dark matter at redshift $z = 26$. Brighter colours correspond to regions of higher concentrations of matter. The blue background image shows the small-scale structure in the top cube (cube size = $[3 \text{ co-moving kpc}]^3$) which has a filamentary topology similar to the large-scale structure in the CDM Universe. The first red image zooms by a factor of one hundred into the average-density high-resolution region. This region was initially a cube of 60 co-moving pc^3 resolved with 64 million particles with a gravitational softening of 10^{-2} co-moving parsecs and masses $1.2 \times 10^{-10} M_{\odot} \equiv M_{\text{moon}}/300$. The final image shows a close-up of one of the individual dark-matter haloes in this region, again zooming in by a factor of one hundred so that the box has a physical length of 0.024 pc. This tiny triaxial Earth-mass halo has a central cusp-like density profile and is smooth, devoid of the substructure that is found within galactic and cluster-mass dark-matter haloes. Even though the index of the power spectrum is very steep on these scales ($n \approx -3$), we find that haloes can collapse before merging into a larger system, rather than the naive expectation that all scales are collapsing simultaneously and thus erasing such structures.

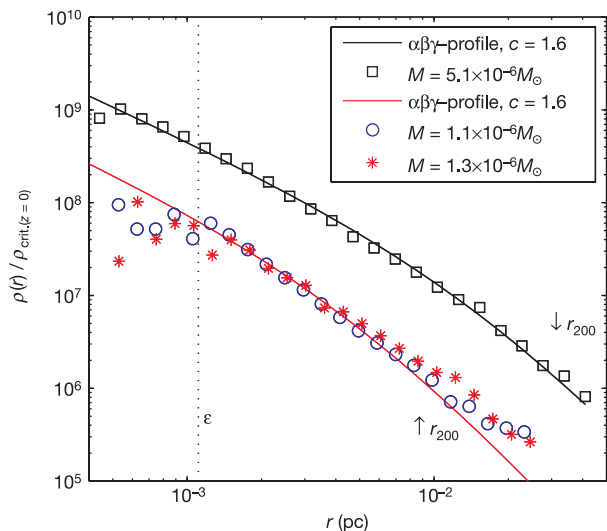


Figure 2 Radial density profiles of three typical minihaloes at redshift $z = 26$. The radial distance is plotted in physical units and we show low-concentration $\alpha\beta\gamma$ -profiles for comparison. We use the mean dark-matter profile inferred from the highest-resolution galaxy-cluster simulations³⁰, that is, $(\alpha\beta\gamma) = (1, 3, 1.2)$. The vertical dotted line indicates our force resolution and the arrows indicate the radii where the halo density is of 200 times the background density. Across the entire range of halo masses from $10^{-6}M_{\odot}$ to $10^1 M_{\odot}$, we find small concentration parameters $c < 3$. We do not observe a trend of concentration with mass, possibly because the haloes all form at a similar epoch, as expected when the power spectrum is so steep.

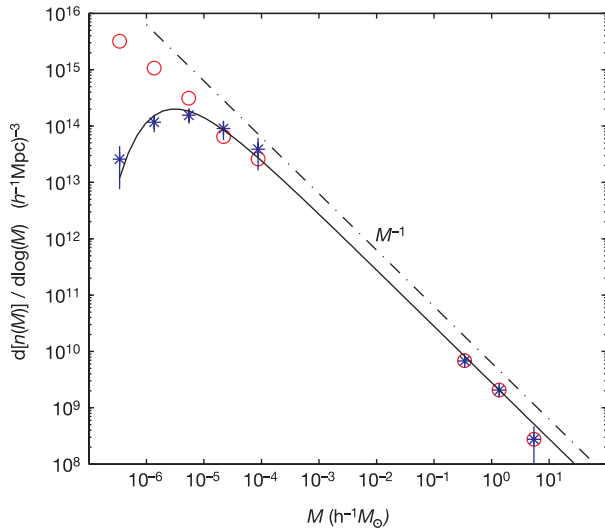


Figure 3 The abundance of collapsed and virialized dark-matter haloes of a given mass. The same region was simulated twice using different types of initial fluctuations: (1) SUSY-CDM with a 100-GeV neutralino (star symbols with 1- σ error bars); and (2) an additional model with no small-scale cut-off to the power spectrum (open circles) as might be produced by an axion dark-matter candidate. Densities are given in co-moving units. Model 2 has a steep mass function down to the resolution limit, whereas run 1 has many fewer haloes below a mass of about $5 \times 10^{-6} h^{-1} M_{\odot} = 3.5 \times 10^{-6} M_{\odot}$, where $h = 0.71$ is the normalized present-day Hubble expansion rate. (Our simulations do not probe the mass range from about $3 \times 10^{-4} h^{-1} M_{\odot}$ to $2 \times 10^{-1} h^{-1} M_{\odot}$). The dashed-dotted line shows an extrapolation of the number density of Galaxy haloes (from ref. 21) assuming $d\ln(M)/d\log M \propto M^{-1}$. The solid line is the function $d\ln(M)/d\log M = 2.8 \times 10^9 (M/h^{-1} M_{\odot})^{-1} \exp[-(M/M_{\text{cut-off}})^{-2/3}]$ in units of $(h^{-1} \text{Mpc})^{-3}$, with a cut-off mass $M_{\text{cut-off}} = 5.7 \times 10^{-6} h^{-1} M_{\odot}$. The power spectrum cut-off is $P(k) \propto \exp[-(k/k_{\text{fs}})^2]$, where k_{fs} is the free streaming scale and assuming $k \propto M^{-1/3}$ motivates the exponent of $-2/3$ in our fitting function.

For a $10^{-6} M_{\odot}$ object $r_E \approx 10^{-7}$ pc, which is much smaller than the size of the mini-haloes considered here, so it is unlikely that gravitational lensing can provide a constraint on their presence, either in our halo or on cosmological path lengths to distant quasars.

Indirect detection is a more interesting possibility and several ongoing and planned experiments aim to detect the atmospheric Cerenkov light from γ -rays produced by neutralino annihilation in the cores of dark haloes^{7,25–28}. Simple scaling arguments show that minihaloes can have high relative luminosities in γ -rays. The absolute γ -ray luminosity of a dark-matter halo with a Navarro–Frenk–White (NFW) density profile is $L \propto \rho_s^2 r_s^3$, where r_s is the scale radius of the NFW profile and $\rho_s = \rho(r_s)$ (ref. 29). The relative luminosity that would arrive at the detector from a halo at a distance d is then $L_{\text{rel}} \propto L d^{-2}$.

Now we compare the relative luminosity of a minihalo at a distance of 0.1 pc (their expected mean separation) to the signal from the centre of the Draco dwarf galaxy:

$$\frac{L_{\text{rel,mini}}}{L_{\text{rel,draco}}} \propto \left(\frac{7 \times 10^6 \rho_{\text{crit}}}{1.7 \times 10^5 \rho_{\text{crit}}} \right)^2 \left(\frac{0.005 \text{ pc}}{300 \text{ pc}} \right)^3 \left(\frac{0.1 \text{ pc}}{82,000 \text{ pc}} \right)^{-1} \approx 5 \quad (2)$$

where we used the typical minihalo properties from our simulations and where ρ_{crit} is the critical energy density. The large abundance of the smallest subclumps compensates their smaller absolute luminosity and the closest of them will be bright sources of γ -rays. The background flux will be enhanced by a boost factor of over two orders of magnitude over a smooth Galactic dark-

matter potential. Current indirect detection experiments such as VERITAS²⁶, HESS²⁷, MAGIC²⁸ or CANGAROO-III²⁵ can probe part of the parameter space predicted by SUSY theory by observing the galactic centre. However this region is dynamically complex because it contains numerous confusing astrophysical γ -ray sources and a supermassive black hole that can erase the central cusp. CDM minihaloes are potentially bright and will not suffer from these problems. All-sky surveys could detect some nearby minihaloes that would have a characteristic extent on the sky, similar to the extent expected for a more distant satellite galaxy like Draco. $\square$

Received 7 September; accepted 3 December 2004; doi:10.1038/nature03270.

1. Peebles, P. J. E. Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *Astrophys. J.* **263**, L1–L5 (1982).
2. Hofmann, S., Schwarz, D. J. & Stöcker, H. Damping scales of neutralino cold dark matter. *Phys. Rev. D* **64**, 083507 (2001).
3. Berezhinsky, V., Dokuchaev, V. & Eroshenko, Y. Small-scale clumps in the galactic halo and dark matter annihilation. *Phys. Rev. D* **68**, 103003 (2003).
4. Green, A. M., Hofmann, S. & Schwarz, D. J. The power spectrum of SUSY-CDM on sub-galactic scales. *Mon. Not. R. Astron. Soc.* **353**, L23–L27 (2004).
5. Jungman, G., Kamionkowski, M. & Griest, K. Supersymmetric dark matter. *Phys. Rep.* **267**, 195–373 (1996).
6. Ellis, J. R., Olive, K. A. & Santoso, Y. Constraining supersymmetry. *New J. Phys.* **4**, 32 (2002).
7. Bertone, G., Hooper, D. & Silk, J. Particle dark matter: Evidence, candidates and constraints. Preprint at (<http://arXiv.org/astro-ph/0404175>), (2004).
8. Spergel, D. N. et al. First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
9. Riess, A. G. et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009–1038 (1998).
10. Perlmutter, S. et al. Measurements of omega and lambda from 42 high-redshift supernovae. *Astrophys. J.* **517**, 565–586 (1999).
11. Tegmark, M. et al. Cosmological parameters from SDSS and WMAP. *Phys. Rev. D* **69**, 103501 (2004).
12. Peebles, P. J. E. Dark matter and the origin of galaxies and globular star clusters. *Astrophys. J.* **277**, 470–477 (1984).
13. Tegmark, M. et al. How small were the first cosmological objects? *Astrophys. J.* **474**, 1–12 (1997).
14. Turner, M. The case for omega mass = 0.33 + i – 0.035. *Astrophys. J.* **576**, L101–L104 (2002).
15. Lake, G. Detectability of gamma-rays from clumps of dark matter. *Nature* **346**, 39–40 (1990).
16. Bergstrom, L., Edsjo, J., Gondolo, P. & Ullio, P. Clumpy neutralino dark matter. *Phys. Rev. D* **59**, 043506 (1999).
17. Calcáneo-Roldán, C. & Moore, B. Surface brightness of dark matter: Unique signatures of neutralino annihilation in the galactic halo. *Phys. Rev. D* **62**, 123005 (2000).
18. Prada, F., Klypin, A., Flix, J., Martinez, M. & Simonneau, E. Astrophysical inputs on the SUSY dark matter annihilation detectability. Preprint at (<http://arXiv.org/astro-ph/0401512>), (2004).
19. Bertschinger, E. Multiscale gaussian random fields at their application to cosmological simulations. *Astrophys. J. Suppl.* **137**, 1–20 (2001).
20. Tasitsiomi, A., Kravtsov, A. V., Gottlober, S. & Klypin, A. A. Density profiles of LCDM clusters. *Astrophys. J.* **607**, 125–139 (2004).
21. Reed, D. et al. Evolution of the mass function of dark matter haloes. *Mon. Not. R. Astron. Soc.* **346**, 565–572 (2003).
22. Diemand, J., Moore, B. & Stadel, J. Velocity and spatial biases in cold dark matter subhalo distributions. *Mon. Not. R. Astron. Soc.* **352**, 535–546 (2004).
23. Moore, B. et al. Dark matter in Draco and the Local Group: Implications for direct detection experiments. *Phys. Rev. D* **64**, 063508 (2001).
24. Schmidt, R. & Wambsganss, J. Limits on MACHOs from microlensing in the double quasar Q0957+561. *Astron. Astrophys.* **335**, 379–387 (1998).
25. Mori, M. et al. (The CANGAROO Collaboration). Status of the CANGAROO-III project. *AIP Conf. Proc.* **558**, 578–581 (2001).
26. Cogan, P. et al. (The VERITAS Collaboration). An overview of the VERITAS prototype telescope and camera. Preprint at (<http://arXiv.org/astro-ph/0408155>), (2004).
27. Hinton, J. A. et al. (The HESS Collaboration). The status of the HESS project. *New Astron. Rev.* **48**, 331–337 (2004).
28. Cortina, J. et al. (The MAGIC Collaboration). Status and first results of the MAGIC telescope. Preprint at (<http://arXiv.org/astro-ph/0407475>), (2004).
29. Kouhiappas, S. M., Zentner, A. R. & Walker, T. P. Observability of gamma rays from neutralino annihilations in the Milky Way substructure. *Phys. Rev. D* **69**, 043501 (2004).
30. Diemand, J., Moore, B. & Stadel, J. Convergence and scatter of cluster density profiles. *Mon. Not. R. Astron. Soc.* **353**, 624–632 (2004).

Acknowledgements We thank A. Green, D. Schwarz, P. Jetzer, M. Miranda, A. Maccio and G. Bertone for discussions. All computations were performed on the zBox supercomputer at the University of Zurich. This work was supported by the Swiss National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.M. (moore@physik.unizh.ch).

Self-similarity of complex networks

Chaoming Song¹, Shlomo Havlin² & Hernán A. Makse¹

¹Levich Institute and Physics Department, City College of New York, New York, New York 10031, USA

²Minerva Center and Department of Physics, Bar-Ilan University, Ramat Gan 52900, Israel

Complex networks have been studied extensively owing to their relevance to many real systems such as the world-wide web, the Internet, energy landscapes and biological and social networks^{1–5}. A large number of real networks are referred to as ‘scale-free’ because they show a power-law distribution of the number of links per node^{1,6,7}. However, it is widely believed that complex networks are not invariant or self-similar under a length-scale transformation. This conclusion originates from the ‘small-world’ property of these networks, which implies that the number of nodes increases exponentially with the ‘diameter’ of the network^{8–11}, rather than the power-law relation expected for a self-similar structure. Here we analyse a variety of real complex networks and find that, on the contrary, they consist of self-repeating patterns on all length scales. This result is achieved by the application of a renormalization procedure that coarse-grains the system into boxes containing nodes within a given ‘size’. We identify a power-law relation between the number of boxes needed to cover the network and the size of the box, defining a finite self-similar exponent. These fundamental properties help to explain the scale-free nature of complex networks and suggest a common self-organization dynamics.

Two fundamental properties of real complex networks have attracted much attention recently: the small-world and the scale-free properties. Many naturally occurring networks are ‘small world’ because we can reach a given node from another one, following the path with the smallest number of links between the nodes, in a very small number of steps. This corresponds to the so-called ‘six degrees of separation’ in social networks¹⁰. It is mathematically expressed by the slow (logarithmic) increase of the average diameter of the network, $\bar{\ell}$, with the total number of nodes N , $\bar{\ell} \approx \ln N$, where ℓ is the shortest distance between two nodes and defines the distance metric in complex networks^{6,8,9,11}. Equivalently, we obtain:

$$N \approx e^{\bar{\ell}/\ell_0} \quad (1)$$

where ℓ_0 is a characteristic length.

A second fundamental property in the study of complex networks arises with the discovery that the probability distribution of the number of links per node, $P(k)$ (also known as the degree distribution), can be represented by a power-law (‘scale-free’) with a degree exponent γ that is usually in the range $2 < \gamma < 3$ (ref. 6):

$$P(k) \approx k^{-\gamma} \quad (2)$$

These discoveries have been confirmed in many empirical studies of diverse networks^{1–4,6,7}.

With the aim of providing a deeper understanding of the underlying mechanism that leads to these common features, we need to probe the patterns within the network structure in more detail. The question of connectivity between groups of interconnected nodes on different length scales has received less attention. But many examples exhibit the importance of collective behaviour, such as interactions between communities within social networks, links between clusters of websites of similar subjects, and the highly modular manner in which molecules interact to keep a cell alive. Here we show that real complex networks, such as the world-wide web (WWW), social, protein–protein interaction networks (PIN) and cellular networks are invariant or self-similar under a length-scale transformation.

This result comes as a surprise, because the exponential increase in equation (1) has led to the general understanding that complex networks are not self-similar, since self-similarity requires a power-law relation between N and ℓ .

How can we reconcile the exponential increase in equation (1) with self-similarity, or (in other words) an underlying length-scale-invariant topology? At the root of the self-similar properties that we unravel in this study is a scale-invariant renormalization procedure that we show to be valid for dissimilar complex networks.

To demonstrate this concept we first consider a self-similar

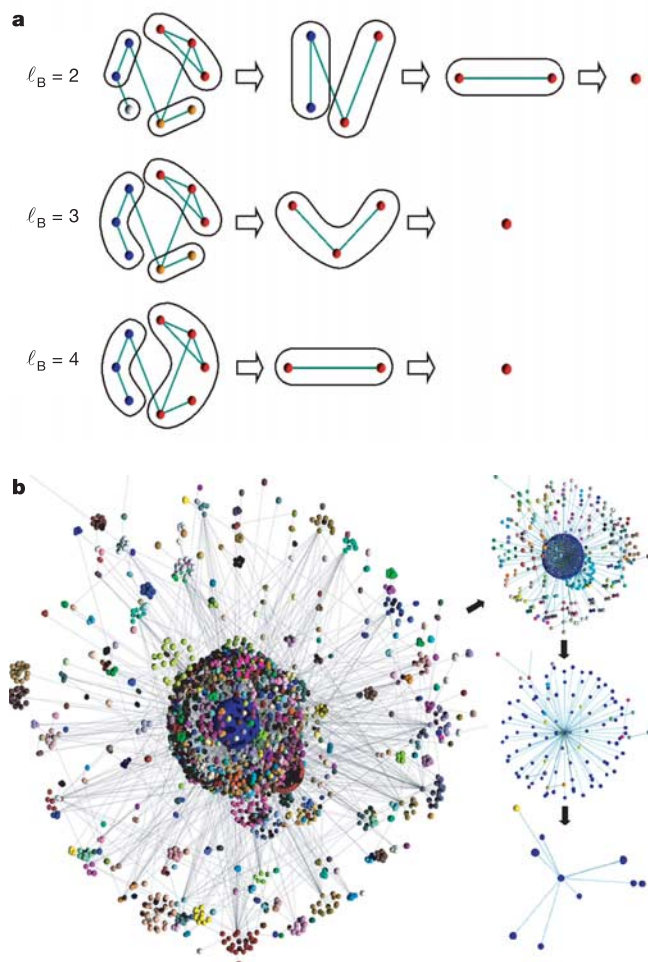


Figure 1 The renormalization procedure applied to complex networks. **a**, Demonstration of the method for different ℓ_B . The first column depicts the original network. We tile the system with boxes of size ℓ_B (different colours correspond to different boxes). All nodes in a box are connected by a minimum distance smaller than the given ℓ_B . For instance, in the case of $\ell_B = 2$, we identify four boxes that contain the nodes depicted with colours red, orange, white and blue, each containing 3, 2, 1 and 2 nodes, respectively. Then we replace each box by a single node; two renormalized nodes are connected if there is at least one link between the unrenormalized boxes. Thus we obtain the network shown in the second column. The resulting number of boxes needed to tile the network, $N_B(\ell_B)$, is plotted in Fig. 2 versus ℓ_B to obtain d_B as in equation (3). The renormalization procedure is applied again and repeated until the network is reduced to a single node (third and fourth columns for different ℓ_B). **b**, The stages in the renormalization scheme applied to the entire WWW. We fix the box size to $\ell_B = 3$ and apply the renormalization for four stages. This corresponds, for instance, to the sequence for the network demonstration depicted in the second row in panel **a**. We colour the nodes in the web according to the boxes to which they belong. The network is invariant under this renormalization, as explained in the legend of Fig. 2d and the Supplementary Information.

network embedded in euclidean space, of which a classical example would be a fractal percolation cluster at criticality¹². To unfold the self-similar properties of such clusters we calculate the fractal dimension using a ‘box-counting’ method and a ‘cluster-growing’ method¹³.

In the first method we cover the percolation cluster with N_B boxes of linear size ℓ_B . The fractal dimension or box dimension d_B is then given by¹⁴:

$$N_B \approx \ell_B^{-d_B} \quad (3)$$

In the second method, the network is not covered with boxes. Instead one seed node is chosen at random and a cluster of nodes centred at the seed and separated by a minimum distance ℓ is calculated. The procedure is then repeated by choosing many seed nodes at random and the average ‘mass’ of the resulting clusters ($\langle M_c \rangle$, defined as the number of nodes in the cluster) is calculated as a function of ℓ to obtain the following scaling:

$$\langle M_c \rangle \approx \ell^{d_f} \quad (4)$$

defining the fractal cluster dimension d_f ¹⁴. Comparing equations (4) and (1) implies that $d_f = \infty$ for complex small-world networks.

For a homogeneous network characterized by a narrow degree distribution (such as a fractal percolation cluster) the box-counting method of equation (3) and the cluster-growing method of equation (4) are equivalent, because every node typically has the

same number of links or neighbours. Equation (4) can then be derived from equation (3) and $d_B = d_f$, and this relation has been regularly used.

The crux of the matter is to understand how we can calculate a self-similar exponent (analogous to the fractal dimension in euclidean space) in complex inhomogeneous networks with a broad degree distribution such as equation (2). Under these conditions equation (3) and (4) are not equivalent, as will be shown below. The application of the proper covering procedure in the box-counting method (equation (3)) for complex networks unveils a set of self-similar properties such as a finite self-similar exponent and a new set of critical exponents for the scale-invariant topology.

Figure 1a illustrates the box-covering method using a schematic network composed of eight nodes. For each value of the box size ℓ_B , we search for the number of boxes needed to tile the entire network such that each box contains nodes separated by a distance $\ell < \ell_B$.

This procedure is applied to several different real networks: (1) a part of the WWW composed of 325,729 web pages that are connected if there is a URL link from one page to another⁶ (<http://www.nd.edu/~networks>); (2) a social network where the nodes are 392,340 actors linked if they were cast together in at least one film¹⁵; (3) the biological networks of protein–protein interactions found in *Escherichia coli* (429 proteins) and *Homo sapiens* (946 proteins) linked if there is a physical binding between them (database available via the Database of Interacting Proteins^{16,17}, other PINs are discussed in the Supplementary Information), and

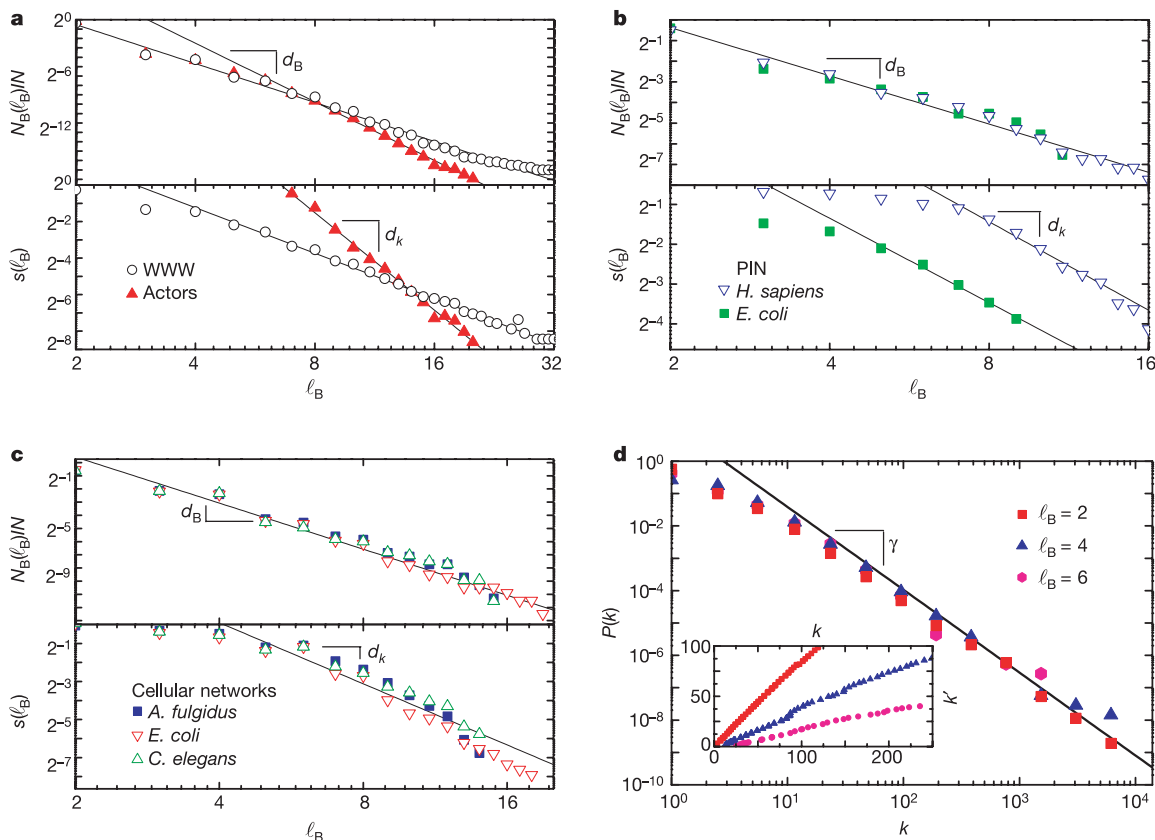


Figure 2 Self-similar scaling in complex networks. **a**, The upper panel shows a log-log plot of N_B versus ℓ_B , revealing the self-similarity of the WWW and actor network according to equation (3). The lower panel shows the scaling of $s(\ell_B)$ versus ℓ_B according to equation (9). The error bars are of the order of the symbol size. **b**, Same as **a** but for two PINs: *H. sapiens* and *E. coli*. Results are analogous to **b** but with different scaling exponents. **c**, Same as **a** for the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans*. **d**, Invariance of the degree distribution of the WWW under the renormalization for different

box sizes, ℓ_B . We show the data collapse of the degree distributions, demonstrating the self-similarity at different scales. The inset shows the scaling of $k' = s(\ell_B)k$ for different ℓ_B , whence we obtain the scaling factor $s(\ell_B)$. Moreover, we also apply the renormalization for a fixed box size, for instance $\ell_B = 3$ as shown in Fig. 1b for the WWW, until the network is reduced to a few nodes, and find that $P(k)$ is invariant under these multiple renormalizations as well, for several iterations (see Supplementary Information).

(4) the cellular networks compiled by ref. 18 using a graph-theoretical representation of all the biochemical pathways based on the WIT integrated-pathway genome database¹⁹ (<http://igweb.integratedgenomics.com/IGwit>) of 43 species from Archaea, Bacteria and Eukarya. Here we show the results for *Archaeoglobus fulgidus*, *E. coli* and *Caenorhabditis elegans*¹⁸; the full database is analysed in the Supplementary Information. It has been previously determined that the WWW and actor networks are small-world and scale-free, characterized by equation (2) with $\gamma = 2.6$ and 2.2, respectively¹. For the PINs of *E. coli* and *H. sapiens* we find $\gamma = 2.2$ and 2.1, respectively. All cellular networks are scale-free with average exponent $\gamma = 2.2$ (ref. 18). We confirm these values and show the results for the WWW in Fig. 2.

Figure 2a and b shows the results of $N_B(\ell_B)$ according to equation (3). They reveal the existence of self-similarity in the WWW, actors and *E. coli* and *H. sapiens* PINs with self-similar exponents $d_B = 4.1$, $d_B = 6.3$, and $d_B = 2.3$ and $d_B = 2.3$, respectively. The cellular networks are shown in Fig. 2c and have $d_B = 3.5$.

We now elaborate on the apparent contradiction between the two definitions of self-similar exponents in complex networks. After performing a renormalization at a given ℓ_B , we calculate the mean mass of the boxes covering the network, $\langle M_B(\ell_B) \rangle$, to obtain:

$$\langle M_B(\ell_B) \rangle \equiv N/N_B(\ell_B) \approx \ell_B^{d_B} \quad (5)$$

which is corroborated by direct measurements for all the networks, and shown in Fig. 3a for the WWW.

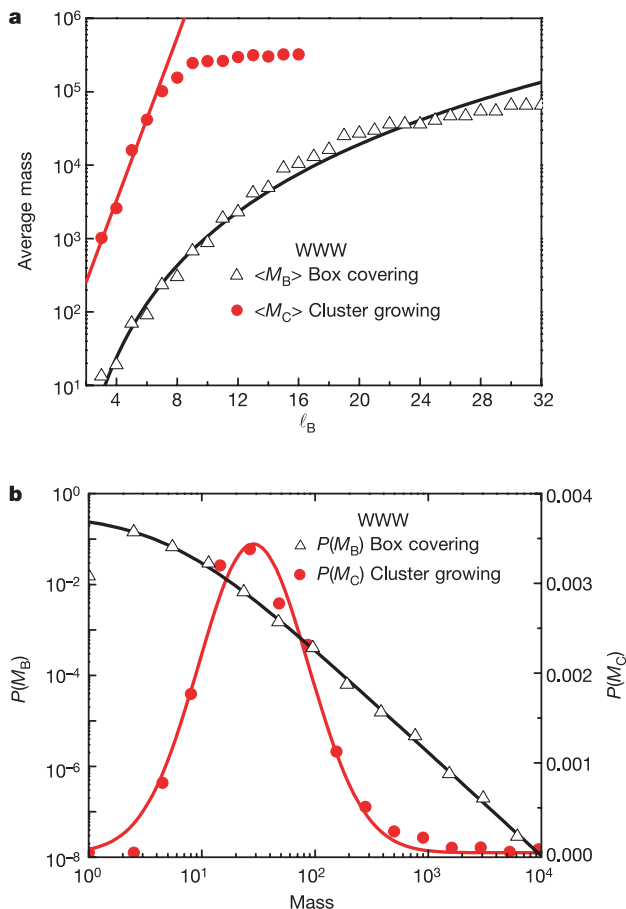


Figure 3 Different averaging techniques lead to qualitatively different results. **a**, Mean value of the box mass in the box-counting method, $\langle M_B \rangle$, and the cluster mass in the cluster growing method, $\langle M_C \rangle$, for the WWW. The solid lines represent the power-law fit for $\langle M_B \rangle$ and the exponential fit for $\langle M_C \rangle$ according to equations (5) and (6), respectively. **b**, Probability distribution of M_B and M_C for $\ell_B = 4$ for the WWW. The curves are fitted by a power-law and a log-normal distribution, respectively.

On the other hand, the average obtained from the cluster-growing method (for this calculation we average over single boxes without tiling the system) gives rise to an exponential growth of the mass:

$$\langle M_C(\ell_B) \rangle \approx \ell_B^{\ell_1} \quad (6)$$

with $\ell_1 \approx 0.78$ in accordance with the small-world effect equation (1), as seen in Fig. 3a.

The topology of scale-free networks is dominated by several highly connected hubs—the nodes with the largest degree—implying that most of the nodes are connected to the hubs via one or very few steps. Therefore the average obtained in the cluster-growing method is biased; the hubs are overrepresented in equation (6) because almost every node is a neighbour of a hub. By choosing the seed of the clusters at random, there is a very large probability of including the hubs in the clusters. On the other hand, the box-covering method is a global tiling of the system, providing a flat average over all the nodes: that is, each part of the network is covered with an equal probability. Once a hub (or any node) is covered, it cannot be covered again. We conclude that equations (3) and (4) are not equivalent for inhomogeneous networks with topologies dominated by hubs with a large degree.

The biased sampling of the randomly chosen nodes is clearly demonstrated in Fig. 3b. We find that the probability distribution of the mass of the boxes for a given ℓ_B is very broad and can be approximated by a power-law: $P(M_B) \approx M_B^{-2.2}$ in the case of the WWW and $\ell_B = 4$. On the other hand, the probability distribution of M_C is very narrow and can be fitted by a log-normal distribution (see Fig. 3b). In the box-covering method there are many boxes with very large and very small masses, in contrast to the peaked distribution in the cluster-growing method, thus showing the biased nature of the latter method in inhomogeneous networks. This biased average leads to the exponential growth of the mass in equation (6) and it also explains why the average distance is logarithmic with N , as in equation (1).

The box-counting method provides a powerful tool for further investigations of network properties because it enables a renormalization procedure, revealing that the self-similar properties and the scale-free degree distribution persist irrespective of the amount of coarse-graining of the network.

Subsequent to the first step of assigning the nodes to the boxes we create a new renormalized network by replacing each box by a single node. Two boxes are then connected, provided that there was at least one link between their constituent nodes. The second column of the panels in Fig. 1a shows this step in the renormalization procedure for the schematic network, while Fig. 1b shows the results for the same procedure applied to the entire WWW for $\ell_B = 3$.

The renormalized network gives rise to a new probability distribution of links, $P(k')$, which is invariant under the renormalization:

$$P(k) \rightarrow P(k') \approx (k')^{-\gamma} \quad (7)$$

Figure 2d supports the validity of this scale transformation by showing a data collapse of all distributions with the same γ according to equation (7) for the WWW.

Further insight arises from relating the scale-invariant properties (equation (3)) to the scale-free degree distribution (equation (2)). Plotting (see inset in Fig. 2d for the WWW) the number of links k' of each node in the renormalized network versus the maximum number of links k in each box of the unrenormalized network exhibits a scaling law:

$$k \rightarrow k' = s(\ell_B)k \quad (8)$$

This equation defines the scaling transformation in the connectivity distribution. Empirically we find that the scaling factor s (< 1) scales with ℓ_B with a new exponent d_k :

$$s(\ell_B) \approx \ell_B^{-d_k} \quad (9)$$

as shown in Fig. 2a for the WWW and actor networks (with $d_k = 2.5$ and $d_k = 5.3$, respectively), in Fig. 2b for the protein networks ($d_k = 2.1$ for *E. coli* and $d_k = 2.2$ for *H. sapiens*) and in Fig. 2c for the cellular networks with $d_k = 3.2$.

Equations (8) and (9) shed light on how families of hierarchical sizes are linked together. The larger the families, the fewer links exist. Surprisingly, the same power-law relation exists for large and small families represented by equation (2).

From equation (7) we obtain $n(k)dk = n'(k')dk'$, where $n(k) = NP(k)$ is the number of nodes with links k and $n'(k') = N'P(k')$ is the number of nodes with links k' after the renormalization (N' is the total number of nodes in the renormalized network). Using equation (8), we obtain $n(k) = s^{1-\gamma}n'(k)$. Then, upon renormalizing a network with N total nodes we obtain a smaller number of nodes N' according to $N' = s^{\gamma-1}N$. The total number of nodes in the renormalized network is the number of boxes needed to cover the unrenormalized network at any given ℓ_B , so we have $N' = N_B(\ell_B)$. Then, from equations (3) and (9) we obtain the relation between the three indexes:

$$\gamma = 1 + d_B/d_k \quad (10)$$

Equation (10) is confirmed for all the networks analysed here (see Supplementary Information). In all cases the calculation of d_B and d_k and equation (10) gives rise to the same γ exponent as that obtained in the direct calculation of the degree distribution. The significance of this result is that the scale-free properties characterized by γ can be related to a more fundamental length-scale invariant property, characterized by the two new indexes d_B and d_k . □

Received 4 August; accepted 30 November 2004; doi:10.1038/nature03248.

1. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
2. Dorogovtsev, S. N. & Mendes, J. F. F. *Evolution of Networks: From Biological Nets to the Internet and the WWW* (Oxford Univ. Press, Oxford, 2003).
3. Pastor-Satorras, R. & Vespignani, A. *Evolution and Structure of the Internet: a Statistical Physics Approach* (Cambridge Univ. Press, Cambridge, 2004).
4. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).
5. Amaral, L. A. N. & Ottino, J. M. Complex networks—augmenting the framework for the study of complex systems. *Eur. Phys. J. B* **38**, 147–162 (2004).
6. Albert, R., Jeong, H. & Barabási, A.-L. Diameter of the World Wide Web. *Nature* **401**, 130–131 (1999).
7. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the Internet topology. *Comput. Commun. Rev.* **29**, 251–262 (1999).
8. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
9. Bollobás, B. *Random Graphs* (Academic, London, 1985).
10. Milgram, S. The small-world problem. *Psychol. Today* **2**, 60–67 (1967).
11. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
12. Bunde, A. & Havlin, S. *Fractals and Disordered Systems* Ch. 2 (eds Bunde, A. & Havlin, S.) 2nd edn (Springer, Heidelberg, 1996).
13. Vicsek, T. *Fractal Growth Phenomena* 2nd edn, Part IV (World Scientific, Singapore, 1992).
14. Feder, J. *Fractals* (Plenum, New York, 1988).
15. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
16. Xenarios, I. et al. DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
17. Database of Interacting Proteins (DIP) (<http://dip.doe-mbi.ucla.edu>) (2000).
18. Jeong, H., Tombor, B., Albert, R., Olvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
19. Overbeek, R. et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acid Res.* **28**, 123–125 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to J. Bruijć for many discussions. This work is supported by the National Science Foundation, Materials Theory. S.H. thanks the Israel Science Foundation and ONR for support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.A.M. (makse@mailaps.org).

Strong polarization enhancement in asymmetric three-component ferroelectric superlattices

Ho Nyung Lee, Hans M. Christen, Matthew F. Chisholm, Christopher M. Rouleau & Douglas H. Lowndes

Condensed Matter Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Theoretical predictions—motivated by recent advances in epitaxial engineering—indicate a wealth of complex behaviour arising in superlattices of perovskite-type metal oxides. These include the enhancement of polarization by strain^{1,2} and the possibility of asymmetric properties in three-component superlattices³. Here we fabricate superlattices consisting of barium titanate (BaTiO₃), strontium titanate (SrTiO₃) and calcium titanate (CaTiO₃) with atomic-scale control by high-pressure pulsed laser deposition on conducting, atomically flat strontium ruthenate (SrRuO₃) layers. The strain in BaTiO₃ layers is fully maintained as long as the BaTiO₃ thickness does not exceed the combined thicknesses of the CaTiO₃ and SrTiO₃ layers. By preserving full strain and combining heterointerfacial couplings, we find an overall 50% enhancement of the superlattice global polarization with respect to similarly grown pure BaTiO₃, despite the fact that half the layers in the superlattice are nominally non-ferroelectric. We further show that even superlattices containing only single-unit-cell layers of BaTiO₃ in a paraelectric matrix remain ferroelectric. Our data reveal that the specific interface structure and local asymmetries play an unexpected role in the polarization enhancement.

Oxide heterostructures with atomically abrupt interfaces, defined by atomically flat surface terraces and single-unit-cell steps, can now be grown on well-prepared single-stepped substrates^{4–7}. This advance has encouraged theoretical investigations that have led to predictions of new artificial materials^{1–3,8–10}. The atomic-scale control of the combining of dissimilar materials is expected to produce striking property enhancements as well as new combinations of desired properties. Here we discuss the experimental realization of one of these predictions, the strain enhancement of ferroelectric polarization. The challenge associated with fabricating such strained structures—the deliberate and controlled deposition of up to hundreds of individual layers—remains a formidable task, for which the principal technique used has been high-vacuum molecular beam epitaxy^{5,11}. However, many insulators do not yield the correct oxide stoichiometry (or expected resulting physical properties) when grown by molecular beam epitaxy. Furthermore, a shortage of electrically conducting oxide substrates and our still-limited understanding of the stability and growth mechanisms of conducting-film electrodes have hindered the electrical characterization of oxide superlattices.

To address these challenges, we have recently shown that atomically flat, electrically conducting SrRuO₃ electrodes can be grown with a surface quality that mimics that of the substrate (Fig. 1a)⁷. Pulsed laser deposition (PLD) has long been regarded as an effective method for synthesizing various oxide heterostructures^{12–15}, but obtaining atomically sharp interfaces has been difficult in the comparatively high-pressure processes needed to maintain oxygen stoichiometry. Here we demonstrate the growth by a high-pressure PLD technique of hundreds of individual perovskite layers of BaTiO₃, SrTiO₃ and CaTiO₃. These superlattices were grown with layer-by-layer control, yielding as-grown samples with compositionally abrupt interfaces, atomically smooth surfaces, and excellent ferroelectric behaviour that indicated oxygen stoichiometry.

Furthermore, whereas several earlier studies have used two-component BaTiO₃/SrTiO₃ structures^{16–19}, we focus here on three-component superlattices (TCSs). This breaks the inversion symmetry^{3,20} that persists with most two-component superlattices, and provides additional freedom in tuning the average lattice parameter by using materials with smaller (CaTiO₃) and larger (BaTiO₃) unit-cell volumes than that of the substrate.

The PLD growth temperature was ~700 °C, with a relatively high oxygen pressure of 10 mtorr to ensure adequate oxidation. Growth conditions were optimized to require about 200 laser pulses per unit cell; this comparatively low deposition rate²¹ allows us to terminate each layer precisely and thus minimize compositional intermixing²². Here we denote unit cells of CaTiO₃, SrTiO₃ and BaTiO₃ as C, S and B, respectively, and define a repeated stacking of *z* unit cells of CaTiO₃ on *y* unit cells of BaTiO₃ on *x* unit cells of SrTiO₃ as S_xB_yC_z, that is, starting the labelling from the bottom layer. The TCS structures studied include S_xB_xC_x (*x* = 1, 2, 5 and 10), S_xB_yC_x (*x* = 2 and *y* = 4, 6, 8), S_xB_yC_y (*x* = 4 and *y* = 2) and S_xB_xC_y (*x* = 2 and *y* = 4).

Figure 1 illustrates the film quality that results from the optimization of the deposition parameters. Cross-sectional atomic-number (Z)-contrast scanning transmission electron microscopy (Z-STEM) shows the compositional abruptness between the SrRuO₃ bottom electrode and a subsequently deposited BaTiO₃ film (Fig. 1a). In this image, a SrO termination of SrRuO₃ can be observed (solid arrow), consistent with RHEED (reflection high-energy electron diffraction) investigations by others²³. Figure 1b depicts an atomic force microscopy (AFM) image of a SrRuO₃ film on a SrTiO₃ substrate. The single terrace steps (~0.4 nm in height) observed on this SrRuO₃ surface are a direct replica of those observed on the SrTiO₃ substrates, and thus provide an ideal starting condition for the TCS growth. In fact, we usually observed pronounced oscillations of the RHEED specular spot to persist even when growing superlattices up to 1 μm in thickness (data not

shown). As shown in Fig. 1c, an AFM image of the top surface of a 200-nm-thick superlattice (S₂B₂C₂) exhibits a surface quality almost identical to that of the substrate and the SrRuO₃ film, again with single-unit-cell steps (~0.4 nm). A cross-sectional Z-contrast image of the corresponding structure in Fig. 1d indeed demonstrates the compositional abruptness of the layers. X-ray scans reveal that the entire structure grows ‘coherently’, that is, with a single value of the in-plane lattice parameter throughout the entire thickness. The long-range periodicity of this layering is further confirmed by X-ray diffraction (XRD): Fig. 1e shows an XRD θ –2 θ scan of a (S₂B₆C₂)₅₀ superlattice grown on SrRuO₃(10 nm)/SrTiO₃, which exhibits clear superlattice peaks associated with the periodicity of the structure. Most structures also show a good peak separation between the Cu K α ₁ and K α ₂ lines (see also the XRD reciprocal space maps below), corresponding to well-defined lattice planes with high crystallinity.

To correlate strain and ferroelectric properties, XRD reciprocal space mapping (RSM) is used to quantify the strain state. Figure 2a shows an area scan through the 114 BaTiO₃ peak in reciprocal space of a 200-nm-thick BaTiO₃ film on a SrRuO₃(4 nm)/SrTiO₃ substrate. Peaks with identical Q_x (in-plane reciprocal lattice unit (1/*d*)) values correspond to crystalline material with identical in-plane lattice parameter. Thus the data of Fig. 2a show that the in-plane lattice parameter of the BaTiO₃ single film is larger than that of the substrate by about 2.5%; that is, the interfacial strain resulting from the substrate is largely relaxed (relieved by the formation of defects).

In contrast, interleaving BaTiO₃ layers with SrTiO₃ and CaTiO₃ layers in a superlattice structure can result in a material in which the in-plane lattice parameter is identical to that of the substrate (and thus smaller than in bulk BaTiO₃). Figure 2b shows the RSM for a S₁B₁C₁ and a S₂B₂C₂ structure; perfect in-plane matching in S_xB_xC_x structures was observed even at values of *x* = 10. Such structures are particularly interesting, as they allow us to experimentally test

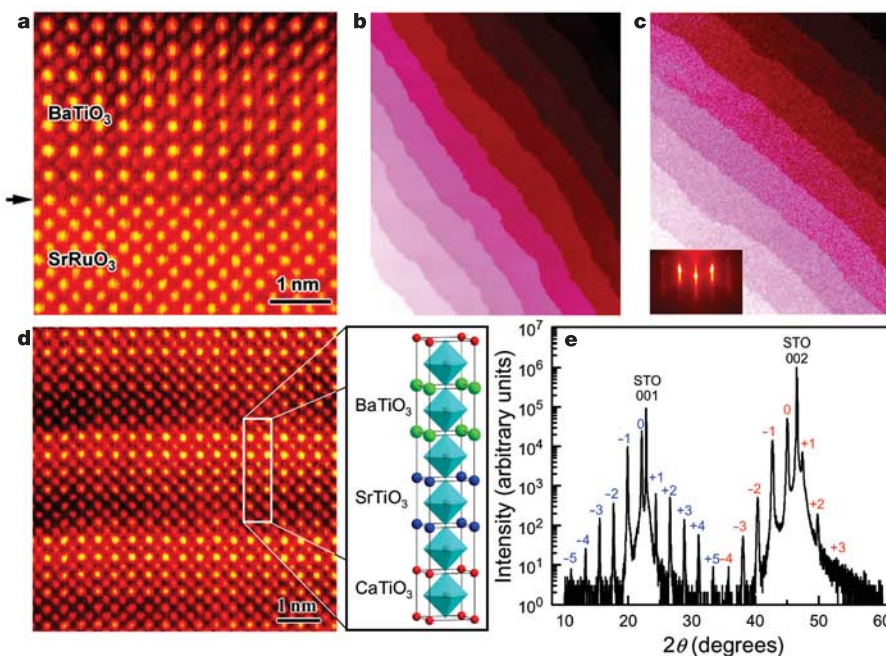


Figure 1 Atomic-scale flatness and compositional abruptness in surfaces and interfaces of conducting SrRuO₃ and artificial superlattices. **a**, Cross-sectional Z-contrast image of interface between BaTiO₃ and SrRuO₃ films, indicated by the black arrow. **b, c**, AFM topographic images (image size: 4 × 5 μm²) with single terrace steps (~0.4 nm) of SrRuO₃ (**b**) and of a 200-nm-thick S₂B₂C₂ superlattice (**c**). Inset in **c**, a RHEED pattern confirming the exceptionally smooth surface. **d**, Cross-sectional Z-contrast image of

compositionally abrupt interfaces in S₂B₂C₂; the diagram shows its atomic structure. (The light-blue octahedra, and the red, green and blue spheres represent TiO₆, and Ca, Ba and Sr, respectively.) **e**, XRD θ –2 θ scan of a S₂B₆C₂ superlattice, confirming the long-range periodicity and high crystallinity. The STO 001 and 002 peaks are from the SrTiO₃ substrate, and the first and second sets of superlattice peaks are marked with blue and red numbers, respectively.

results from calculations¹. These calculations indicate a strain enhancement of the ferroelectric polarization in BaTiO₃ by about 50% as a consequence of an increase in the tetragonality (resulting from a decrease of the in-plane lattice parameter). Competing with this strain enhancement is the effect of the ‘dilution’ of the ferroelectric material in a paraelectric matrix. Density functional calculations on SrTiO₃/BaTiO₃ superlattices¹ indicate that the polarization is significantly larger than what would be expected from the SrTiO₃ and BaTiO₃ volume fractions alone, owing to electric-field-induced polarization in SrTiO₃. Calculations further predict ferroelectricity in superlattices containing only single-unit-cell layers of ferroelectric BaTiO₃ in an otherwise paraelectric matrix. In addition, even for structures containing single-unit-cell layers of BaTiO₃, the numerical results agree remarkably well with those from macroscopic considerations of independent ferroelectric and paraelectric slabs with dielectric constants ϵ_{FE} and ϵ_{PE} , respectively. In this electrostatic model, the average polarization is found to be

$$P_{\text{avg}} = P_{\text{FE}}[1 + (t_{\text{PE}}/t_{\text{FE}})(\epsilon_{\text{FE}}/\epsilon_{\text{PE}})]^{-1} \quad (1)$$

where P_{FE} is the polarization of the ferroelectric layer, and t_{FE} and t_{PE} are the thicknesses of ferroelectric and paraelectric layers, respectively.

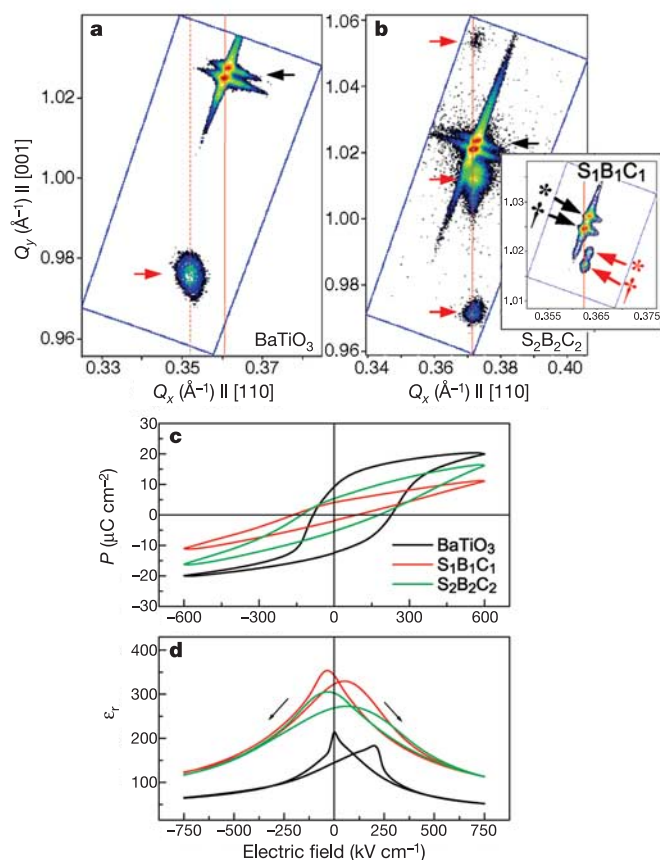


Figure 2 Strain states of a BaTiO₃ single film and superlattices on SrRuO₃/SrTiO₃ substrates and their ferroelectric properties. **a, b**, XRD-RSMs of a BaTiO₃ thin film and of a S₂B₂C₂ superlattice, respectively (Q_x and Q_y in units of $1/d$). A RSM of S₁B₁C₁ is also shown in the inset of **b** (same axes as main figure), showing clearly peaks from the Cu $K\alpha_1$ ($\dagger$) and $K\alpha_2$ ($\ast$) lines. The solid red vertical lines mark the 114 SrTiO₃ Cu $K\alpha_1$ peak position, and the dotted red line that of BaTiO₃. Peaks marked with black and red arrows correspond respectively to reflections from the SrTiO₃ substrate, and from the film or superlattice layer and satellites. **c, d**, $P(E)$ at 100 Hz (**c**) and $\epsilon(E)$ hysteresis loops at 100 kHz (**d**), from 200-nm-thick BaTiO₃ (black curves), S₁B₁C₁ (red) and S₂B₂C₂ (green).

For our TCSs containing one ferroelectric and two paraelectric components (SrTiO₃ with thickness t_{STO} and dielectric constant ϵ_{STO} , and CaTiO₃ with t_{CTO} and ϵ_{CTO}), the electrostatic considerations (equation (1)) are easily generalized by letting $t_{\text{PE}} = t_{\text{STO}} + t_{\text{CTO}}$ and $\epsilon_{\text{PE}} = \epsilon_{\text{STO}}\epsilon_{\text{CTO}}(t_{\text{STO}} + t_{\text{CTO}})/(\epsilon_{\text{STO}}t_{\text{CTO}} + \epsilon_{\text{CTO}}t_{\text{STO}})$. Therefore, on the basis of electrostatics alone, the S₁B₁C₁ and S₂B₂C₂ structures should exhibit identical values of the polarization, whereas in fact the remanent polarization $P_r(\text{S}_1\text{B}_1\text{C}_1) \approx 3.0 \mu\text{C cm}^{-2}$ and $P_r(\text{S}_2\text{B}_2\text{C}_2) \approx 5.4 \mu\text{C cm}^{-2}$ (Fig. 2c), indicating that the interfaces have a significant effect. Moreover, as expected from previous work on SrTiO₃/BaTiO₃ superlattices^{17,19}, the dielectric constants (ϵ) as a function of electric field (E) as determined from the capacitance–voltage $C(V)$ curves at 100 kHz (Fig. 2d) are larger than that of the BaTiO₃ film.

Increasing the number of neighbouring Ba layers results in an increase of the enhancement of the polarization, pointing to the limitations of simple electrostatic considerations (equation (1)). The TiO₆ octahedra (which in the simplest picture are most responsible for the ferroelectric polarization) can be grouped into two classes (see below): those bound on both sides by the same A-site cations (for example, Ba-surrounded TiO₆) and those forming the interface between BaTiO₃, SrTiO₃ or CaTiO₃ (interfacial TiO₆).

To clarify the interplay between strain enhancement of the polarization and interfacial effects, we modified the number of interfaces and of neighbouring BaTiO₃ layers by interleaving identical SrTiO₃ and CaTiO₃ stackings with different numbers of BaTiO₃ layers (while keeping the total thickness fixed). Figure 3a–c shows RSMs (near the 114 SrTiO₃ reflection) of the S₂B₄C₂, S₂B₆C₂ and S₂B₈C₂ superlattices, respectively. A coherent in-plane lattice parameter is obtained only if the BaTiO₃ thickness in each superlattice period does not exceed the combined thickness of the SrTiO₃ and CaTiO₃ layers; consequently, a partial strain relaxation of 0.05% (S₂B₆C₂) and 1.21% (S₂B₈C₂) is observed in Fig. 3b and c, respectively. The evolution of the in-plane lattice parameter is shown in Fig. 4c as a function of the stacking sequence. In Fig. 4d, we show the polarization for the same structures. Clearly, the strongest polarization enhancement is achieved by the proper balancing of two competing requirements: the BaTiO₃ layers must be thick enough to contain a sufficient amount of non-interfacial TiO₆ octahedra (compare Fig. 4e), but thin enough to remain fully strained. This produces a maximum polarization for the S₂B₄C₂ structure (Fig. 4a). Despite the low volume fraction of 50%

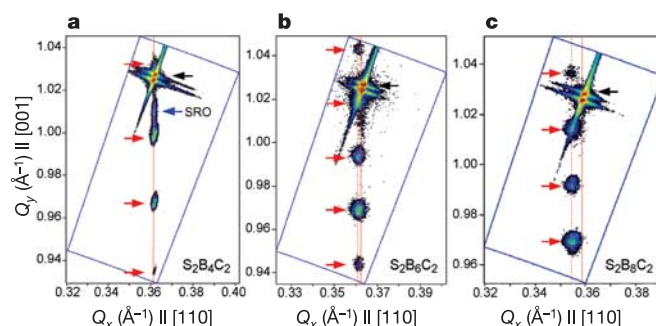


Figure 3 Evolution of the in-plane strain with changes in the thickness of BaTiO₃ layers for fixed thickness of SrTiO₃ and CaTiO₃. **a–c**, XRD-RSMs of S₂B₄C₂ (**a**), S₂B₆C₂ (**b**) and S₂B₈C₂ (**c**) superlattices, respectively, showing a gradual increase in the in-plane lattice parameter. (Solid and dotted red vertical lines mark the peak(s) of the substrate and superlattice, respectively.) The RSMs are recorded around the 114 SrTiO₃ reflection. Peaks marked with black and red arrows correspond respectively to reflections from the SrTiO₃ substrate, and those from the superlattice layer and satellites. The peak labelled ‘SRO’ in **a** (indicated by the blue arrow) corresponds to the 114 reflection of the relatively thick SrRuO₃ film (16 nm).

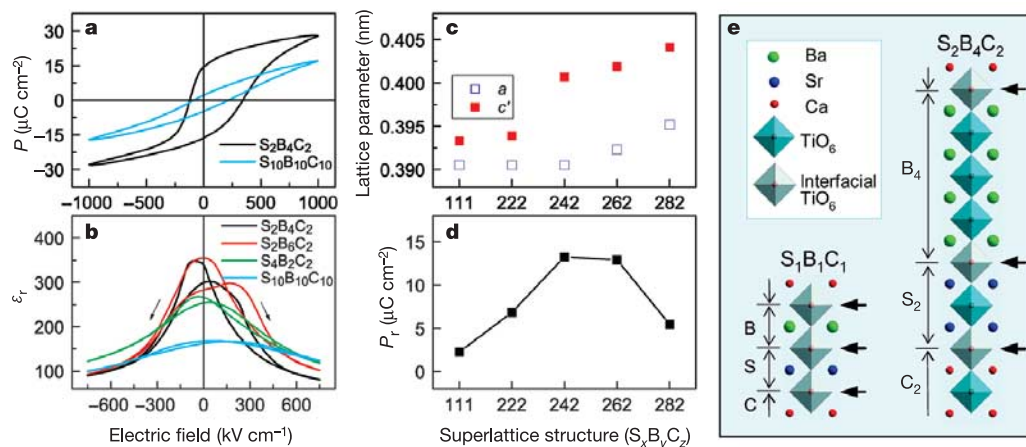


Figure 4 Polarization enhancement, changes in asymmetry, and evolution of strain in TCS structures. **a**, $P(E)$ curves of $\text{S}_2\text{B}_4\text{C}_2$ ($P_r \approx 16.5 \mu\text{C cm}^{-2}$) and $\text{S}_{10}\text{B}_{10}\text{C}_{10}$ ($P_r \approx 3.5 \mu\text{C cm}^{-2}$) at $1,000 \text{ kV cm}^{-1}$. (Note that this electric field is higher than that applied to measure the polarization of the BaTiO_3 film, whose polarization is near saturation at 400 kV cm^{-1} , whereas the $\text{S}_2\text{B}_4\text{C}_2$ requires a higher electric field ($\sim 700 \text{ kV cm}^{-1}$)). **b**, $\epsilon(E)$ curves of $\text{S}_2\text{B}_4\text{C}_2$ (black line), $\text{S}_2\text{B}_6\text{C}_2$ (red line), $\text{S}_4\text{B}_2\text{C}_2$ (green line) and $\text{S}_{10}\text{B}_{10}\text{C}_{10}$ (light blue line) showing differing degrees of asymmetry. **c**, In-plane (open squares) and out-of-plane (filled squares) lattice parameters of various

superlattices. c' corresponds to the supercell c divided by the number of constituent perovskites in a supercell. **d**, P_r values from $E = \pm 750 \text{ kV cm}^{-1}$ loops. The partially relaxed $\text{S}_2\text{B}_6\text{C}_2$ structure was measured with $E = \pm 650 \text{ kV cm}^{-1}$ because of its lower breakdown strength. **e**, Diagrams of supercells showing the different local environments possible for the TiO_6 -octahedra (bound by the same or different A-site cations). Heterointerfacial TiO_6 octahedra are shaded in grey and indicated by solid black arrows.

BaTiO_3 , we observe an almost 50% increase in polarization ($P_r \approx 16.5 \mu\text{C cm}^{-2}$ at 1 MV cm^{-1}) over that of the BaTiO_3 sample. This polarization is significantly larger than the value of $10.9 \mu\text{C cm}^{-2}$ that would be predicted from electrostatic considerations alone, using bulk values for the effective dielectric constants $\epsilon_{\text{STO}} = 300$, $\epsilon_{\text{CTO}} = 186$ and $\epsilon_{\text{BTO}} = 160$.

The fundamental difference between interfacial cells of BaTiO_3 and TiO_6 -octahedra in an environment resembling that of bulk BaTiO_3 is further illustrated by considering structures of equal eight-unit-cell periodicity but different arrangements, that is, $\text{S}_4\text{B}_2\text{C}_2$, $\text{S}_2\text{B}_4\text{C}_2$ and $\text{S}_2\text{B}_6\text{C}_4$. All these structures are fully strained, and as pointed out above, the $\text{S}_2\text{B}_4\text{C}_2$ structure shows very strong polarization enhancement. In contrast, the other two samples of this series are characterized by $P_r < 3 \mu\text{C cm}^{-2}$ (data not shown), slightly less than what would be expected from our BaTiO_3 data and the lower BaTiO_3 volume fraction (equation (1)).

The fact that equation (1) is insufficient to describe our data for three-component structures but almost perfectly reproduces the calculated behaviour of two-component $\text{BaTiO}_3/\text{SrTiO}_3$ super0 lattices points to additional effects that are operating here. One key weakness of applying equation (1) to the TCS configuration is the treatment of the $\text{CaTiO}_3/\text{SrTiO}_3$ bilayer as a simple dielectric with a passive interface. In reality, local distortions at the $\text{SrTiO}_3/\text{CaTiO}_3$ interface may also play a part, possibly even inducing a polarization in the neighbouring SrTiO_3 cells. Furthermore, for the samples considered here, the environment of the BaTiO_3 layers is strongly asymmetric. Considering the theoretically predicted enhancement of the polarization due to breaking of inversion symmetry will therefore be important in understanding these materials³. A previous report on similar TCS structures observed strong indications of asymmetry in $\epsilon(E)$, but without ferroelectric hysteresis¹¹. In contrast, the samples studied here exhibit strong ferroelectric switching with hysteresis curves both in $P(E)$ and $\epsilon(E)$. However, asymmetry is observed here even in simple BaTiO_3 films. Therefore, we cannot neglect the possible effect of the dissimilar top and bottom electrodes (Pt and SrRuO_3 , respectively) on the asymmetric behaviour. Nevertheless, close inspection of the $\epsilon(E)$ curves in Fig. 4b shows a varying degree of asymmetry within each individual peak for different stacking sequences. This behaviour is not easily explained simply as a consequence of different electrode

materials, but is expected to result from compositionally broken inversion symmetry³.

This analysis of atomic-scale precision TCS structures thus leads to two observations. First, the use of three-component materials, as dictated by the need for near-perfect lattice matching to the substrate, reveals the subtle effects of a non-inversion-symmetric environment and the role of a large number of coupled heterointerfaces. Second, these structures have allowed the experimental verification of theoretical predictions of strain-enhanced ferroelectric polarization of BaTiO_3 , even in structures containing a relatively small volume fraction of BaTiO_3 . These results, together with the advance in pulsed laser deposition reported here, illustrate how enhancements of material properties—needed for applications in sensors, high- k dielectrics, and piezoelectric and ferroelectric devices—can be produced. □

Methods

Before growing the TCSs, epitaxial SrRuO_3 films (4–16 nm in thickness) were grown by PLD on single-stepped (001) SrTiO_3 substrates using a KrF excimer laser (wavelength $\lambda = 248 \text{ nm}$) at 700°C in 100 mtorr O_2 . The TCSs were also grown by PLD at $\sim 700^\circ\text{C}$ in 10 mtorr O_2 using sintered stoichiometric CaTiO_3 and BaTiO_3 targets, and a single-crystal SrTiO_3 target. We have previously shown that this combination of pressure and temperature lies within the stability range for our SrRuO_3 films⁷. In addition, the use of CaTiO_3 in the superlattice enhances the compressive strain, allows the average in-plane lattice parameter of the superlattice to match that of the substrate, and induces a wider range of elastic behaviour at the interface with BaTiO_3 . The laser repetition rate and laser fluence were kept at 10 Hz and 2 J cm^{-2} , respectively. The deposition of each individual unit cell was monitored via the intensity oscillations of the specular spot in RHEED pattern. The topmost layer was capped by one or two unit cells of SrTiO_3 and the final thickness of all superlattices was fixed at 200 nm . For the electrical characterization, sputtered Pt top electrodes ($100 \mu\text{m}$ in diameter) were used together with the conducting SrRuO_3 bottom electrode.

Received 17 September; accepted 3 December 2004; doi:10.1038/nature03261.

- Neaton, J. B. & Rabe, K. M. Theory of polarization enhancement in epitaxial $\text{BaTiO}_3/\text{SrTiO}_3$ superlattices. *Appl. Phys. Lett.* **82**, 1586–1588 (2003).
- Pertsev, N. A., Tagantsev, A. K. & Setter, N. Phase transitions and strain-induced ferroelectricity in SrTiO_3 epitaxial thin films. *Phys. Rev. B* **61**, R825–R829 (2000).
- Sai, N., Meyer, B. & Vanderbilt, D. Compositional inversion symmetry breaking in ferroelectric perovskites. *Phys. Rev. Lett.* **84**, 5636–5639 (2000).
- Kawasaki, M. et al. Atomic control of the SrTiO_3 crystal surface. *Science* **266**, 1540–1542 (1994).
- Schlom, D. G. et al. Oxide nano-engineering using MBE. *Mater. Sci. Eng. B* **87**, 282–291 (2001).
- Yamada, H. et al. Engineered interface of magnetic oxides. *Science* **305**, 646–648 (2004).
- Lee, H. N., Christen, H. M., Chisholm, M. F., Rouleau, C. M. & Lowndes, D. H. Thermal stability of epitaxial SrRuO_3 films as a function of oxygen pressure. *Appl. Phys. Lett.* **84**, 4107–4109 (2004).

8. Okamoto, S. & Millis, A. J. Electronic reconstruction at an interface between a Mott insulator and a band insulator. *Nature* **428**, 630–633 (2004).
9. Junquera, J. & Ghosez, P. Critical thickness for ferroelectricity in perovskite ultrathin films. *Nature* **422**, 506–509 (2003).
10. Glunchuk, M. D. & Morozovska, A. N. The internal electric field originating from the mismatch effect and its influence on ferroelectric thin film properties. *J. Phys. Condens. Matter* **16**, 3517–3531 (2004).
11. Warusawithana, M. P., Colla, E. V., Eckstein, J. N. & Weissman, M. B. Artificial dielectric superlattices with broken inversion symmetry. *Phys. Rev. Lett.* **90**, 36802 (2003).
12. Lee, H. N., Hesse, D., Zakharov, N. & Gösele, U. Ferroelectric $\text{Bi}_{3.25}\text{La}_{0.75}\text{Ti}_3\text{O}_{12}$ films of uniform *a*-axis orientation on silicon substrates. *Science* **296**, 2006–2009 (2002).
13. Christen, H. M. *et al.* An improved continuous compositional-spread technique based on pulsed-laser deposition and applicable to large substrate areas. *Rev. Sci. Instrum.* **74**, 4058–4062 (2003).
14. Zheng, H. *et al.* Multiferroic $\text{BaTiO}_3\text{-CoFe}_2\text{O}_4$ nanostructures. *Science* **303**, 661–663 (2004).
15. Norton, D. P. Synthesis and properties of epitaxial electronic oxide thin-film materials. *Mater. Sci. Eng. R* **43**, 139–247 (2004).
16. Tabata, H., Tanaka, H. & Kawai, T. Formation of artificial $\text{BaTiO}_3/\text{SrTiO}_3$ superlattices using pulsed laser deposition and their dielectric properties. *Appl. Phys. Lett.* **65**, 1970–1972 (1994).
17. O'Neill, D., Bowman, R. M. & Gregg, J. M. Dielectric enhancement and Maxwell-Wagner effects in ferroelectric superlattice structures. *Appl. Phys. Lett.* **77**, 1520–1522 (2000).
18. Shimuta, T. *et al.* Enhancement of remanent polarization in epitaxial $\text{BaTiO}_3/\text{SrTiO}_3$ superlattices with asymmetric structure. *J. Appl. Phys.* **91**, 2290–2294 (2002).
19. Kim, J. *et al.* Large nonlinear dielectric properties of artificial $\text{BaTiO}_3/\text{SrTiO}_3$ superlattices. *Appl. Phys. Lett.* **80**, 3581–3583 (2002).
20. Ogawa, Y. *et al.* Nonlinear magneto-optical Kerr rotation of an oxide superlattice with artificially broken symmetry. *Phys. Rev. Lett.* **90**, 217403 (2003).
21. Müller, D. A., Nakagawa, N., Ohtomo, A., Grazul, J. L. & Hwang, H. Y. Atomic-scale imaging of nanoengineered oxygen vacancy profiles in SrTiO_3 . *Nature* **430**, 657–661 (2004).
22. Lee, H. N. *et al.* Compositionally asymmetric tri-color superlattices grown by pulsed laser deposition. *Mater. Res. Soc. Symp. Proc.* **784**, C3.24 (2004).
23. Rijnders, G., Blank, D. H. A., Choi, J. & Eom, C. B. Enhanced surface diffusion through termination conversion during epitaxial SrRuO_3 growth. *Appl. Phys. Lett.* **84**, 505–507 (2004).

Acknowledgements This work was supported by the US Department of Energy under contract with the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, as part of a BES NSET initiative on Nanoscale Cooperative Phenomena and of the LDRD programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.N.L. (hnlee@ornl.gov).

Another continental pool in the terrestrial silicon cycle

Isabelle Basile-Doelsch^{1,2}, Jean Dominique Meunier² & Claude Parron²

¹IRD-Réunion, UR037, LSTUR, BP 172, 97492 St-Denis Messagerie Cedex, France

²CEREGE UMR CNRS/ Université Paul Cézanne/IRD, Europole de l'Arbois, BP80, 13545 Aix-en-Provence Cedex 4, France

Silicon is the second most abundant element on Earth. It is an important nutrient for phytoplankton¹ and is readily absorbed by terrestrial vegetation²; it also assists the removal of carbon dioxide from the atmosphere through the weathering of silicates³. But the continental cycle of silicon is not well known, and only a few studies have attempted to use silicon stable isotopes (^{28}Si , ^{29}Si and ^{30}Si)^{4–13} to quantify the continental silicon reservoirs. Dissolved silicon in sea and river waters forms a reservoir of mean isotopic value +1.1‰ (refs 7, 10). It is enriched in ^{30}Si with respect to the igneous rocks reservoir, which has a mean isotopic value of −0.3‰ (refs 4, 9). This enrichment can only be produced by a major fractionation during weathering, and should result in the formation of a continental ^{30}Si -depleted reservoir. Such a reservoir, however, has not been identified to date. Here we analyse silicon isotopes of *in situ* quartz from a sandstone series in France, using a new-generation secondary ion mass spectrometry apparatus. We show that quartz that precipitates as siliceous cements forms a strongly ^{30}Si -depleted reservoir with isotopic values down to −5.7‰, a more negative value than

any previously published for terrestrial samples. Our findings suggest that quartz re-precipitation plays an important role in the biogeochemical cycle of silicon.

In the budget of silicon isotopes, the mean $\delta^{30}\text{Si}$ value of igneous rocks is −0.3‰ (refs 4, 9). This reservoir constitutes the main pool of Si that will be weathered in continental environments to reach the rivers and then the sea. However, the river and marine samples measured to date both have an average of +1.1‰ (refs 7, 10). In the continental environment, the cycle of Si should thus take into account this +1.4‰ enrichment found for Si in rivers relative to Si in igneous rocks. Up to now, two processes that fractionate Si isotopes have been suspected: biomineralization of Si by plants into phytoliths^{4,9,13} and precipitation of Si into clays during weathering^{9–13}. However, they do not seem to represent sufficiently depleted ^{30}Si pools to counterbalance the waters' enrichment. A

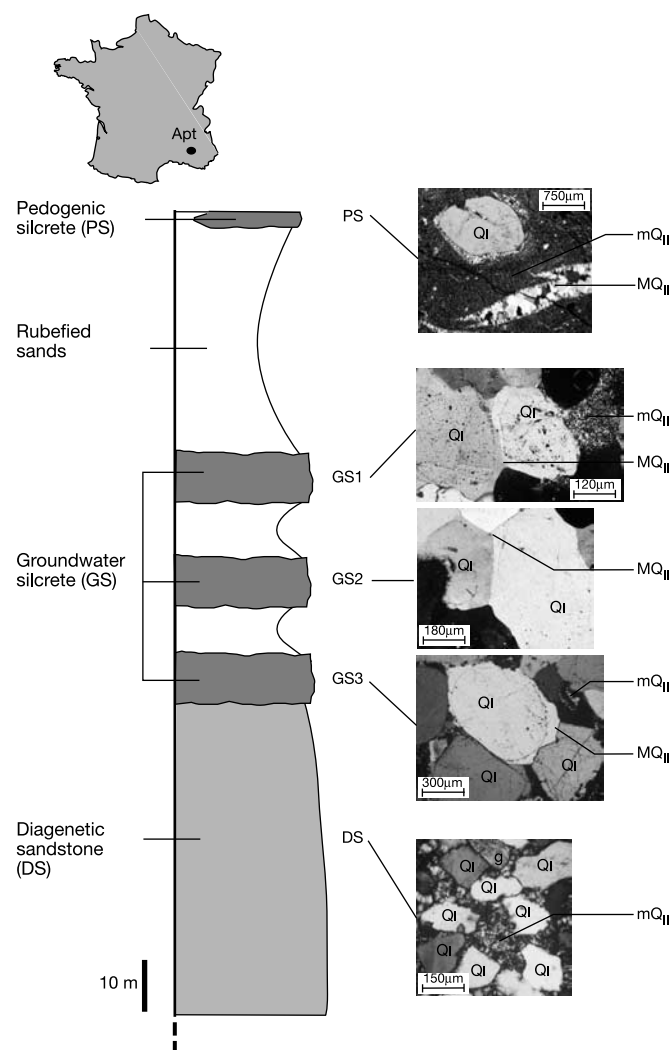


Figure 1 Schematic section of the Apt series, showing five levels of silicification. Four silicites were formed in continental environments (one pedogenic silcrete, PS; three groundwater silicites, GS1, GS2 and GS3). The cement of the basal sandstone formed in a marine sedimentary environment (diagenetic sandstone, DS). In the PS sample, mQt forms the soil matrix while MQ_{II} is found in desiccation cracks. In GS samples, MQ_{II} consists of subeuhedral overgrowths on primary detrital quartz Qt , and mQt is found in pores remaining after glauconite weathering. In the DS sample, glauconite (g) is present and the quartz cement consists only of mQt microcrystals. Rubefied sands, sands coloured red by iron oxides (see Supplementary Information).

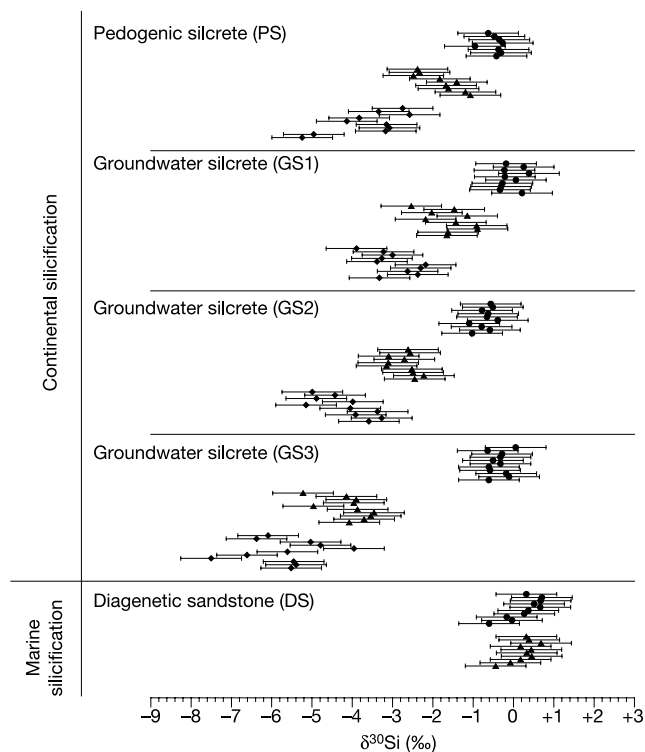


Figure 2 $\delta^{30}\text{Si}$ of quartz from the five levels of silicified rocks. Filled circles, primary detrital quartz (Q_I); filled triangles, microcrystalline quartz (mQ_{II}); and filled diamonds, macrocrystalline quartz (MQ_{II}). Error bars are $\pm 0.75\text{‰}$ (2σ).

third process, silicification, stores Si in continents for millions of years. Silicification is a low-temperature physical-chemical process that cements surficial material such as rocks, sediments, saprolites or soils by various forms of silica. Although silicified rocks have been widely recognized in numerous areas (such as Australia¹⁴, southern¹⁵ and northern Africa¹⁶, parts of western Europe¹⁷, parts of North¹⁸ and South America¹⁹, and the Arabian Gulf²⁰), they have never been considered as potential depleted ^{30}Si pools. The objective of this study is to show that quartz that precipitates as siliceous cements is strongly depleted in ^{30}Si , and may represent an important pool in the budget of silicon.

The studied samples originate from a sandstone series from Apt, southeast France^{21,22} (see Supplementary Information). At the top of the stratigraphic series, sample PS represents a level of pedogenic silicification; the three intermediate levels are groundwater silicifi-

cation samples (GS1, GS2, GS3); and, at the base of the series, sample DS was cemented during marine diagenesis (Fig. 1). Petrographical thin sections (Fig. 1) show that three types of quartz are present in the samples: (1) large rounded grains of primary detrital quartz (which we refer to as Q_I); (2) microcrystalline cementing quartz (mQ_{II}); and (3) macrocrystalline cementing quartz (MQ_{II}). *In situ* analysis of Si isotopes were performed by secondary ion mass spectrometry (SIMS) by focusing the bombarding ions on each individual mineral phase of the sample. Although the precision of this technique is lower than the usually used methods (isotopic ratio mass spectrometry, IRMS, and multi-collection induced coupled plasma mass spectrometry, MC-ICP-MS; see Supplementary Information), it has the advantage of *in situ* single-grain analysis, thus avoiding the extraction procedures which are used on bulk samples (such as total rock, set of diatoms, and water).

The $\delta^{30}\text{Si}$ results are reported in Fig. 2 and Table 1. In the pedogenic silcrete (PS), the Q_I has a mean $\delta^{30}\text{Si}$ value ($\bar{\delta}^{30}\text{Si}$) of $-0.5 \pm 0.4\text{‰}$ (see Methods for definition of reported uncertainties). The mQ_{II} of the soil matrix has more negative values, with a $\bar{\delta}^{30}\text{Si}$ value of $-1.8 \pm 1.0\text{‰}$. The most negative values of PS samples are found for MQ_{II} precipitated in desiccation cracks, with a $\bar{\delta}^{30}\text{Si}$ value of $-3.6 \pm 1.8\text{‰}$. A similar trend for the $\delta^{30}\text{Si}$ values is also observed for the groundwater silicified samples (GS1, GS2, GS3): (1) the cementing quartz (mQ_{II} and MQ_{II}) has systematically more negative values than Q_I , and (2), the $\bar{\delta}^{30}\text{Si}$ of MQ_{II} is systematically more negative than the $\bar{\delta}^{30}\text{Si}$ of mQ_{II} . In contrast, for marine diagenetic silicification (DS), the mQ_{II} have a $\bar{\delta}^{30}\text{Si}$ value of $0.2 \pm 0.6\text{‰}$, comparable to that found in Q_I ($0.3 \pm 0.9\text{‰}$).

The results can be compared with existing $\delta^{30}\text{Si}$ data reported for terrestrial rocks of equivalent origin. Concerning the primary quartz Q_I , our $\delta^{30}\text{Si}$ values range for all samples from $-1.1 \pm 0.75\text{‰}$ to $0.7 \pm 0.75\text{‰}$, with a mean $\bar{\delta}^{30}\text{Si}$ value of $-0.3 \pm 0.4\text{‰}$ (49 measurements). This range is in good agreement with the published $\delta^{30}\text{Si}$ range of igneous rocks (-1‰ to $+0.4\text{‰}$, with a $\bar{\delta}^{30}\text{Si}$ of -0.3‰)^{4,9}, which confirms the reliability of SIMS measurements. The $\delta^{30}\text{Si}$ values found here for quartz cements of silicified rocks in continental environments (mQ_{II} and MQ_{II}) range from -0.9 down to $-7.5 \pm 0.75\text{‰}$ (80 measurements). However, it seems in this case more rigorous to only consider for each type of quartz, not extreme values, but rather the mean values, $\bar{\delta}^{30}\text{Si}$, which range from $-1.6 \pm 1.1\text{‰}$ (mQ_{II} of GS1) down to $-5.7 \pm 1.9\text{‰}$ (MQ_{II} of GS3). The reported values obtained by IRMS on bulk sandstones range from $-0.5 \pm 0.2\text{‰}$ to $+1.1 \pm 0.2\text{‰}$ (8 measurements)⁹. These values are more positive than ours because bulk analyses necessarily include the isotopic composition of a large amount of Q_I types together with a possible low amount of secondary quartz. Our reported $\bar{\delta}^{30}\text{Si}$ values are much more negative than the most depleted sample ever measured, which

Table 1 $\delta^{30}\text{Si}$ values of quartz from silicified rocks

	Type of quartz	Number of analyses	Minimum of $\delta^{30}\text{Si}$ (‰)	Maximum of $\delta^{30}\text{Si}$ (‰)	Mean of $\delta^{30}\text{Si} \pm 2\sigma$ (‰)
Pedogenic silcrete, PS	Q_I	8	-1.0	-0.3	-0.5 ± 0.4
	mQ_{II}	9	-2.5	-1.1	-1.8 ± 1.0
	MQ_{II}	10	-5.2	-2.6	-3.6 ± 1.8
Groundwater silcrete, GS1	Q_I	10	-0.3	+0.4	-0.1 ± 0.5
	mQ_{II}	10	-2.5	-0.9	-1.6 ± 1.1
	MQ_{II}	10	-3.9	-2.2	-3.0 ± 1.1
Groundwater silcrete, GS2	Q_I	10	-1.1	-0.4	-0.7 ± 0.4
	mQ_{II}	10	-4.6	-2.2	-2.9 ± 1.3
	MQ_{II}	10	-5.1	-3.3	-4.2 ± 1.4
Groundwater silcrete, GS3	Q_I	11	-0.6	+0.1	-0.4 ± 0.5
	mQ_{II}	10	-5.2	-3.5	-4.1 ± 1.2
	MQ_{II}	11	-7.5	-4.0	-5.7 ± 1.9
Diagenetic sandstone, DS	Q_I	10	-0.6	+0.7	$+0.3 \pm 0.9$
	mQ_{II}	10	-0.4	+0.7	$+0.2 \pm 0.6$

Data are from the Apt sandstone series (see Fig. 1). Q_I , primary quartz; mQ_{II} , microcrystalline quartz; MQ_{II} , macrocrystalline quartz. Minimum and maximum $\delta^{30}\text{Si}$ values are $\pm 0.75\text{‰}$.

consisted of sponge spicules ($-3.7 \pm 0.2\text{‰}$)^{4,8}. Our results thus increase the reported $\delta^{30}\text{Si}$ range of terrestrial samples to $-5.7\text{‰} < \delta^{30}\text{Si} < 3.4\text{‰}$. Moreover, they show (to our knowledge for the first time) the existence in continental environments of a highly ^{30}Si -depleted pool, which may counterbalance the observed ^{30}Si enrichment of river waters ($+1.1\text{‰}$)^{7,10}.

In order to explain such a fractionation of Si isotopes in the quartz of cements, our results can be compared with known fractionation values. Si isotope fractionation ranges from -0.3‰ to -3.8‰ for chemical^{9,11,12} and biochemical^{5,8,11,12} precipitation of Si (see Supplementary Information). Thus, simple precipitations alone can not be responsible for the most negative values measured in this work. The explanation is given by the physical-chemical processes that control the precipitation of various types of quartz in silcretes. According to Thiry²³, the most soluble minerals²⁴ (here mQ_{II}) precipitate first. At the scale of the porous space, the less soluble minerals (MQ_{II}) precipitate afterwards from solutions whose Si is provided by the dissolution of the most soluble minerals (mQ_{II}). Silicification thus would cumulate Si isotope fractionation at each step of crystallization, with an estimated averaged fractionation $\Delta_{\text{quartz-sol}} \approx -1.5\text{‰}$ (which represents an effective degree of fractionation of about -2.4‰ between solution and precipitate, see Supplementary Information). For groundwater silcretes, additional dissolution/crystallization occurs at the scale of the sedimentary formation. The vertical superposition of silicification levels corresponds to successively deepening levels of the water table²³. Each new silicified level would be fed by water that passes through the unsaturated zone above and dissolves the remaining mQ_{II} of these previously silicified levels (see Supplementary Information). Such mechanisms add supplementary crystallization steps that may explain the very negative values of $\delta^{30}\text{Si}$ of the cements of the GS2 and GS3 samples. From these mechanisms, the fractionation can be

modelled by using the Rayleigh equations (see Supplementary Information). The isotopic value of the output water, which varies from $+0.8\text{‰}$ to $+1.1\text{‰}$, simulates with success the natural values of river waters.

These results show that silicification plays an important role, and should be included in the biogeochemical cycle of Si. We propose a continental cycle with three compartments—rivers and lakes, soils and aquifers (Fig. 3). In each compartment, the different Si pools are taken into account (Si in primary and secondary minerals, biogenic Si and dissolved Si) together with their up-to-date range of $\delta^{30}\text{Si}$ signatures and inter-pool Si transfers²⁴. The pedogenic and groundwater silicification pools are dissociated, because their environmental context is different. Pedogenic silicification is associated with arid climates, and its contribution to the Si cycle probably followed global and/or regional climatic variations. Pedogenic silicification that has been identified in recent soils is probably still active^{19,23}. On the other hand, groundwater silicification is not climate dependent, and may occur in an appropriate aquifer anywhere on Earth²³. These two types of silicification may therefore contribute significantly to the Si cycle (both now and in the past). However, the contribution of silcretes to the Si cycle is difficult to estimate quantitatively.

Assuming that silcretes constitute the unique reservoir of weathered Si on continents, the isotopic budget (see Supplementary Information) shows that $\sim 30\%$ of Si would be locked up in silcretes, the remaining 70% being dissolved in river waters. However, other continental reservoirs (such as phytoliths and clays) have also been identified, but their quantitative contribution to the Si cycle remains unknown. In addition, the rare isotopic data available to date for clays and phytoliths can not clearly explain the isotopic values of rivers, whereas quartz precipitates at low temperature can explain them. The objective of future work on the Si cycle will be the quantification of the respective contribution of each reservoir,

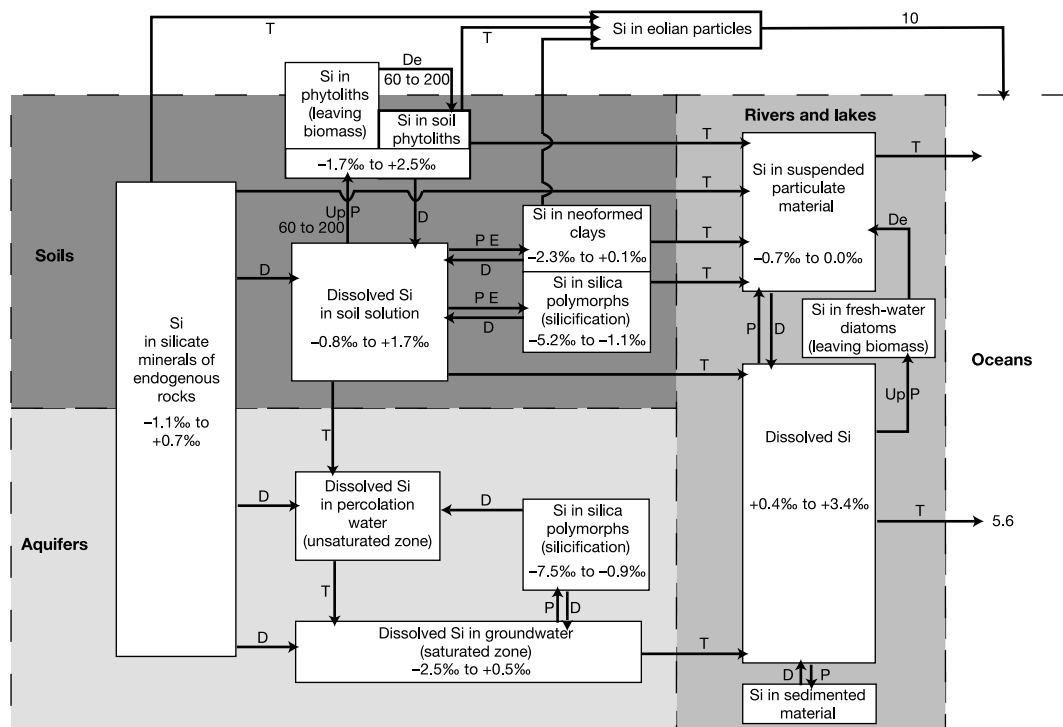


Figure 3 Biogeochemical cycle of Si in continental environments (adapted from ref. 24). D, dissolution; P, precipitation; T, transport; E, epigenesis; Up, uptake; De, death. $\delta^{30}\text{Si}$ ranges are from: endogenous rocks (ref. 4 and this work); soil solution¹²; phytoliths^{4,9,13};

neoformed clays^{7,9,13}; pedogenic and groundwater silicifications (this work); ground water (this work and ref. 13); river water^{7,10}; and suspended particulate material¹⁰. Numbers labelling arrows show fluxes in Tmol (10^{12} mol; refs 29, 30).

keeping in mind that some reservoirs may be still unknown, and also that the mechanism of successive precipitation/dissolution/precipitation, which generates very ^{30}Si -depleted pools, may also occur in systems other than silcretes—such as clay minerals²⁵. □

Methods

Mean river water isotopic value

The mean value of $\delta^{30}\text{Si}$ in rivers represents a mean of seven $\delta^{30}\text{Si}$ measurements of dissolved Si sampled in seven different rivers^{7,10}. When several data were available in a river, we considered the sample closest to the river mouth.

Notation of isotopic data

The $^{30}\text{Si}/^{28}\text{Si}$ ratios are reported in $\delta^{30}\text{Si}$ notation, in ‰ variation relative to the $^{30}\text{Si}/^{28}\text{Si}$ ratio of the NBS28 quartz standard:

$$\delta^{30}\text{Si} = \left(\frac{(^{30}\text{Si}/^{28}\text{Si})_{\text{sample}}}{(^{30}\text{Si}/^{28}\text{Si})_{\text{NBS28}}} - 1 \right) \times 1,000$$

The NBS28 standard is a NIST standard prepared from an African sand from which the fraction 100–177 μm was collected after acid wash. In previous studies of terrestrial samples, all the $^{30}\text{Si}/^{28}\text{Si}$ ratios are normalized to NBS28, except in ref. 4 where $^{30}\text{Si}/^{28}\text{Si}$ ratios are normalized to rose quartz. In this study, as in ref. 9, reported values of ref. 4 are corrected by -0.3‰ (ref. 26).

In the text, single measurements are denoted $\delta^{30}\text{Si}$ and the corresponding uncertainty ($\pm 2\sigma$) is the analytical precision. Means of single measurements performed on a given type of quartz in a sample are denoted $\delta^{30}\text{Si}$, and the corresponding uncertainty ($\pm 2\sigma$) is the dispersion of the single measurements around the mean. The fractionation ($\Delta_{\text{quartz-sol}}$) is the difference between the $\delta^{30}\text{Si}$ of the solid phase and the $\delta^{30}\text{Si}$ of the dissolved Si of the initial solution from which the solid phase precipitates (see Supplementary Information).

SIMS measurements

Up to now and because of the rather low precision of the method, SIMS (see refs 9 and 27 for reviews) and nanoSIMS had only been used on interstellar material where the range of $\delta^{30}\text{Si}$ values is larger than in terrestrial samples. However, things have much improved with the arrival of new-generation SIMS machines, such as the one used for this study, the Cameca 1270. SIMS now represents an interesting alternative method for measuring Si isotopes in Earth samples. The configuration of the SIMS consisted of a Cs^+ primary beam, with a beam size of around 50 μm in diameter. The detection of secondary ions used the L2 Faraday cage of the multicollector system for $^{28}\text{Si}^-$ and the FC2 axial Faraday cage of the monocollector system for $^{30}\text{Si}^-$. The mass resolution ($M/\Delta M$) was $\sim 5,000$, allowing the separation of $^{30}\text{Si}^-$ and $^{29}\text{SiH}^-$ (ref. 28). 15 blocks of two cycles of 3 s were carried out, representing a total sample analysing time of ~ 6 min.

A flat sample surface can be obtained either by mounting individual grains in epoxy resin followed by a polish, or by making polished thin sections (30 μm thick) when samples to be analysed are rocks. Thin sections present the advantage that minerals can be easily identified by classical optical properties. In the present study, five samples of rocks as well as individual grains of quartz standards were prepared as thin sections. Around six grains of an in-house quartz standard were included in each thin section for standard-sample bracketing normalization. The in-house quartz standard is a pure Brazilian optical grade crystal which was used during the Second World War by the navy in sonar transducers. The crystals are untwinned, both optically and electrically, and are probably uniform in composition. An additional thin section was made with grains of the in-house standard and grains of the NBS28 standard. Finally, an additional epoxy resin impregnated sample was prepared with grains of rose quartz and NBS28 quartz to check the accuracy of the method. As instrumental mass fractionation is supposed to affect both the quartz standards and the quartz sample equally, the true isotopic compositions of the quartz measured in the samples can be calculated as follows:

$$\delta^{30}\text{Si}(\text{‰}) = \left(\frac{(^{30}\text{Si}/^{28}\text{Si})_{\text{sample}}^{\text{measured-sp}}}{(^{30}\text{Si}/^{28}\text{Si})_{\text{house-std}}^{\text{measured-sp}}} \times \frac{(^{30}\text{Si}/^{28}\text{Si})_{\text{house-std}}^{\text{measured-st}}}{(^{30}\text{Si}/^{28}\text{Si})_{\text{NBS28}}^{\text{measured-st}}} - 1 \right) \times 1,000$$

where ‘measured-sp’ represents the measurements performed on the thin sections of samples, and ‘measured-st’ the measurements performed on the thin sections of standards. An additional epoxy resin impregnated sample was prepared with grains of rose quartz and NBS28 quartz to check the accuracy of the method. In that case, the $\delta^{30}\text{Si}$ was calculated as follows:

$$\delta^{30}\text{Si}(\text{‰}) = \left(\frac{(^{30}\text{Si}/^{28}\text{Si})_{\text{RoseQuartz}}^{\text{measured-ep}}}{(^{30}\text{Si}/^{28}\text{Si})_{\text{NBS28}}^{\text{measured-ep}}} - 1 \right) \times 1,000$$

where ‘measured-ep’ represents the measurements performed in the epoxy resin.

During the two sessions of measurements, the internal analytical precision measured on the in-house standard (45 measurements) ranged from 0.04‰ to 0.08‰ (1 σ). The reproducibility of the measurement of the in-house standard was $\pm 0.48\text{‰}$ (2 σ , $n=15$) during the first session, and $\pm 0.34\text{‰}$ (2 σ , $n=30$) during the second session. However, we observed a slight drift between the two sessions, which gives a lower standard deviation for the whole data set of $\pm 0.75\text{‰}$ (2 σ , $n=45$). As the error due to reproducibility is much higher than the internal analytical error, the analytical precision is $\pm 0.75\text{‰}$ (2 σ). The rose quartz was measured 11 times during the two sessions to test the accuracy. The mean value of $\delta^{30}\text{Si}$ is $+0.26\text{‰}$ and the standard deviation is $\pm 0.62\text{‰}$ (2 σ , $n=11$).

Received 25 March; accepted 9 November 2004; doi:10.1038/nature03217.

1. Ragueneau, O. *et al.* A review of the Si cycle in the modern ocean: recent progress and missing gaps in the application of biogenic opal as a paleoproductivity proxy. *Glob. Planet. Change* **26**, 317–365 (2000).
2. Alexandre, A., Meunier, J. D., Colin, F. & Koud, J. M. Plant impact on the biogeochemical cycle of silicon and related weathering processes. *Geochim. Cosmochim. Acta* **61**, 677–682 (1997).
3. Berner, R. A. The rise of plants and their effect on weathering and atmospheric CO_2 . *Science* **276**, 544–546 (1997).
4. Douthitt, C. B. The geochemistry of the stable isotopes of silicon. *Geochim. Cosmochim. Acta* **46**, 1449–1458 (1982).
5. De La Rocha, C. L., Breszinski, M. A. & DeNiro, M. J. Fractionation of silicon isotopes by marine diatoms during biogenic silica formation. *Geochim. Cosmochim. Acta* **61**, 5051–5056 (1997).
6. De La Rocha, C. L., Breszinski, M. A., DeNiro, M. J. & Shemesh, A. Silicon isotope composition of diatoms as an indicator of past oceanic change. *Nature* **395**, 680–683 (1998).
7. De La Rocha, C. L., Breszinski, M. A. & DeNiro, M. J. A first look at the distribution of the stable isotopes of silicon in natural waters. *Geochim. Cosmochim. Acta* **64**, 2467–2477 (2000).
8. De La Rocha, C. L. Silicon isotope fractionation by marine sponges and the reconstruction of the silicon isotope composition of ancient deep water. *Geology* **31**, 423–426 (2003).
9. Ding, T. *et al.* *Silicon Isotope Geochemistry* (Geological Publishing House, Beijing, 1996).
10. Ding, T., Wan, D., Wang, C. & Zhang, F. Silicon isotope compositions of dissolved silicon and suspended matter in the Yangtze River, China. *Geochim. Cosmochim. Acta* **68**, 205–216 (2004).
11. Ziegler, K., Chadwick, O. A., Kelly, E. F., Breszinski, M. A. & DeNiro, M. in *Goldschmidt Conference 1135* (Cambridge Publication, Oxford, 2000).
12. Ziegler, K., Chadwick, O. A., Kelly, E. F. & Breszinski, M. A. in *6th Int. Symp. on the Geochemistry of the Earth's Surface* 103–106 (University of Hawaii, Honolulu, 2002).
13. Ziegler, K., Chadwick, O. A., Kelly, E. F. & Breszinski, M. A. *Goldschmidt Conference, Geochim. Cosmochim. Acta* **66**(S1), abstract A881 (2002).
14. Arakel, A. V., Jacobson, G., Salehi, M. & Hill, C. M. Silicification of the calcrete in paleodrainage basins of Australian arid zone. *Aust. J. Earth Sci.* **36**, 73–89 (1989).
15. Nash, D. J. & Shaw, P. A. Silica and carbonate relationships in silcrete-calcrete intergrade duricrusts from the Kalahari of Botswana and Namibia. *J. Afr. Earth Sci.* **27**, 11–25 (1998).
16. Abdel-Wahab, A., Salem, A. M. K. & McBride, E. F. Quartz cement of meteoric origin in silcrete and nonsilcrete sandstones, Lower Carboniferous, western Sinai, Egypt. *J. Afr. Earth Sci.* **27**, 277–290 (1998).
17. Thiry, M. & Simon-Coinçon, R. Tertiary paleoweatherings and silcretes in the southern Paris Basin. *Catena* **26**, 1–26 (1996).
18. Chadwick, O., Hendricks, D. M. & Nettleton, W. D. Silicification of Holocene soils in northern Monitor Valley, Nevada. *Soil Soc. Am.* **53**, 159–164 (1989).
19. Elsass, F., Dubroeuq, D. & Thiry, M. Diagenesis of silica minerals from clay minerals in volcanic soils of Mexico. *Clay Miner.* **35**, 477–489 (2000).
20. Khalaf, F. I. Petrography and diagenesis of silcrete from Kuwait, Arabian Gulf. *J. Sedim. Petrol.* **58**, 1014–1022 (1988).
21. Parron, C., Nahon, D., Fritz, B., Paquet, H. & Millot, G. Désilicification et quartzification par altération des grès albiens du Gard. Modèles géochimiques de la genèse des dalles quartzitiques et silcretes. *Sci. Geol. Bull.* **29**, 273–284 (1976).
22. Parron, C. & Guendon, J. L. Les altérites médiocretacées de Provence (bauxites et ocre): caractères et relations paléogéographiques. *Géol. Fr.* **2**, 137–150 (1985).
23. Thiry, M. in *Soils and Sediments: Mineralogy and Geochemistry* (eds Paquet, H. & Clauer, N.) 191–221 (Springer, Berlin, 1997).
24. Dove, P. M. in *Chemical Weathering Rates of Silica Minerals* (eds White, A. F. & Brantley, S. L.) 235–290 (Reviews in Mineralogy, Mineralogical Society of America, Washington DC, 1996).
25. Girard, J.-P., Freyssinet, P. & Chazot, G. Unraveling climatic changes from intraprofile variation in oxygen and hydrogen isotopic composition of goethite and kaolinite in laterites: an integrated study from Yaou, French Guiana. *Geochim. Cosmochim. Acta* **64**, 409–426 (2000).
26. Molini-Vesko, C., Mayeda, T. K. & Clayton, R. N. Isotopic composition of silicon in meteorites. *Geochim. Cosmochim. Acta* **50**, 2719–2726 (1986).
27. Zinner, E. in *New Frontiers in Stable Isotope Research: Laser Probes, Ion Probes, and Small-sample Analysis* (eds Shanks, W. C. III & Criss, R. E.) 145–162 (Bulletin 1890, USGS, Denver, 1989).
28. McKeegan, K. D., Walker, R. M. & Zinner, E. Ion microprobe isotopic measurements of individual interplanetary dust particles. *Geochim. Cosmochim. Acta* **49**, 1971–1987 (1985).
29. Conley, D. J. Terrestrial ecosystems and the global biogeochemical cycle. *Glob. Biogeochem. Cycles* **16**, doi:10.1029/2002GB001894 (2002).
30. Tréguer, P. *et al.* The silica balance in the world ocean: a reestimate. *Science* **268**, 375–379 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Leblanc for technical support for sample preparation; I. Friedman for providing the quartz standard; M. Chausson for the development of Si isotope measurements on the SIMS Cameca 1270 (CRPG, Nancy, France); M. Champenois, C. Champion and D. Mangin for their assistance during the National INSU Facility Sessions; and A. Alexandre, M. Chausson and W. Stone for improving the manuscript. This work was supported by CEREGE funds and the French national programme “Programme National sur les Sols et l’Erosion”.

Authors’ contributions J.D.M. is the instigator of Si isotope measurements in samples formed in weathering systems; C.P. and I.B.D. performed the field sampling and described thin sections; I.B.D. prepared the thin sections with included standards, performed the analysis, developed the model and wrote the Letter.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.B.-D. (basile@cerege.fr).

Uncertainty in predictions of the climate response to rising levels of greenhouse gases

D. A. Stainforth¹, T. Aina¹, C. Christensen², M. Collins³, N. Faull¹, D. J. Frame¹, J. A. Kettleborough⁴, S. Knight¹, A. Martin², J. M. Murphy³, C. Piani¹, D. Sexton³, L. A. Smith⁵, R. A. Spicer⁶, A. J. Thorpe⁷ & M. R. Allen¹

¹Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK

²Computing Laboratory, University of Oxford, Parks Road, Oxford OX1 3QD, UK

³Hadley Centre for Climate Prediction and Research, Met Office, Exeter EX1 3PB, UK

⁴Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, UK

⁵London School of Economics, London WC2A 2AE, UK

⁶Department of Earth Sciences, The Open University, Milton Keynes MK7 6AA, UK

⁷Department of Meteorology, University of Reading, Reading RG6 6BB, UK

The range of possibilities for future climate evolution^{1–3} needs to be taken into account when planning climate change mitigation and adaptation strategies. This requires ensembles of multi-decadal simulations to assess both chaotic climate variability and model response uncertainty^{4–9}. Statistical estimates of model response uncertainty, based on observations of recent climate change^{10–13}, admit climate sensitivities—defined as the equilibrium response of global mean temperature to doubling levels of atmospheric carbon dioxide—substantially greater than 5 K. But such strong responses are not used in ranges for future climate change¹⁴ because they have not been seen in general circulation models. Here we present results from the ‘climateprediction.net’ experiment, the first multi-thousand-member grand ensemble of simulations using a general circulation model and thereby explicitly resolving regional details^{15–21}. We find model versions as realistic as other state-of-the-art climate models but with climate sensitivities ranging from less than 2 K to more than 11 K. Models with such extreme sensitivities are critical for the study of the full range of possible responses of the climate system to rising greenhouse gas levels, and for assessing the risks associated with specific targets for stabilizing these levels.

As a first step towards a probabilistic climate prediction system we have carried out a grand ensemble (an ensemble of ensembles) exploring uncertainty in a state-of-the-art model. Uncertainty in model response is investigated using a perturbed physics ensemble⁴ in which model parameters are set to alternative values considered plausible by experts in the relevant parameterization schemes⁹. Two or three values are taken for each parameter (see Methods); simulations may have several parameters perturbed from their standard model values simultaneously. For each combination of parameter values (referred to here as a ‘model version’) an initial-condition ensemble²² is used, creating an ensemble of ensembles. Each individual member of this grand ensemble (referred to here as a ‘simulation’) explores the response to changing boundary conditions²² by including a period with doubled CO₂ concentrations.

The general circulation model (GCM) is a version of the Met Office Unified Model consisting of the atmospheric model HadAM3²³, at standard resolution⁹ but with increased numerical stability, coupled to a mixed-layer ocean. This allows us to explore the effects of a wide range of uncertainties in the way the atmosphere is represented, while avoiding a long spin-up for each model version. Each simulation involves three 15-year phases: (1) calibration, to deduce the ocean heat-flux convergence field used in the subsequent phases; (2) control, used to quantify the relevance of the particular model version and heat-flux convergence field; and (3)

doubled CO₂, to explore the response to changing boundary conditions.

Individual simulations are carried out using idle processing capacity on personal computers volunteered by members of the general public¹⁹. This distributed-computing method^{16,18,19} leads to a continually expanding data set of results, requiring us to use a specified subset of data available at a specific point in time. The analysis presented here uses 2,578 simulations (>100,000 simulated years), chosen to explore combinations of perturbations in six parameters.

The 2,578 simulations contain 2,017 unique simulations (duplicates are used to verify the experimental design—see Methods). Figure 1a shows the grand ensemble frequency distribution of global mean, annual mean, near-surface temperature (T_g) in these 2,017 simulations, as it develops through each phase. Some model versions show substantial drifts in the control phase owing to the use of a simplified ocean (see Supplementary Information). We remove unstable simulations (see Methods) and average over initial-condition ensembles of identical model versions to reduce sampling uncertainty. The frequency distribution of initial-condition-ensemble-mean time series of T_g for the resulting 414 model versions (for which the initial-condition ensembles involve 1,148 independent stable simulations) is shown in Fig. 1b. Six of these model versions show a significant cooling tendency in the doubled-CO₂ phase. This cooling is also due to known limitations with the use of a simplified ocean (see Supplementary Information) so these simulations are excluded from the remaining analysis of sensitivity.

The frequency distribution of the simulated climate sensitivities (see Methods) for the remaining model versions is shown in Fig. 2a and ranges from 1.9 to 11.5 K. Two key features are that relatively few model versions have sensitivities less than 2 K, and the long tail of the distribution extending to very high values; 4.2% are >8 K. Most sensitivities cluster round 3.4 K, the value for the unperturbed model, suggesting that many of the parameter combinations

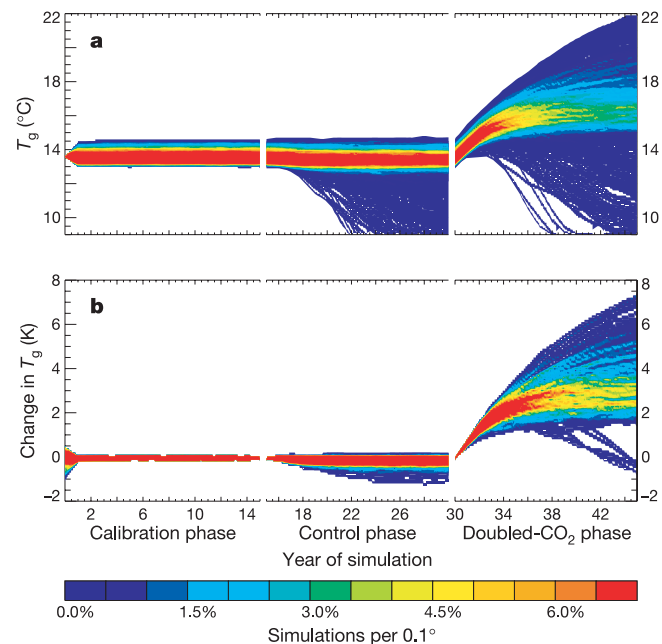


Figure 1 Frequency distributions of T_g (colours indicate density of trajectories per 0.1 K interval) through the three phases of the simulation. **a**, Frequency distribution of the 2,017 distinct independent simulations. **b**, Frequency distribution of the 414 model versions. In **b**, T_g is shown relative to the value at the end of the calibration phase and where initial-condition ensemble members exist, their mean has been taken for each time point.

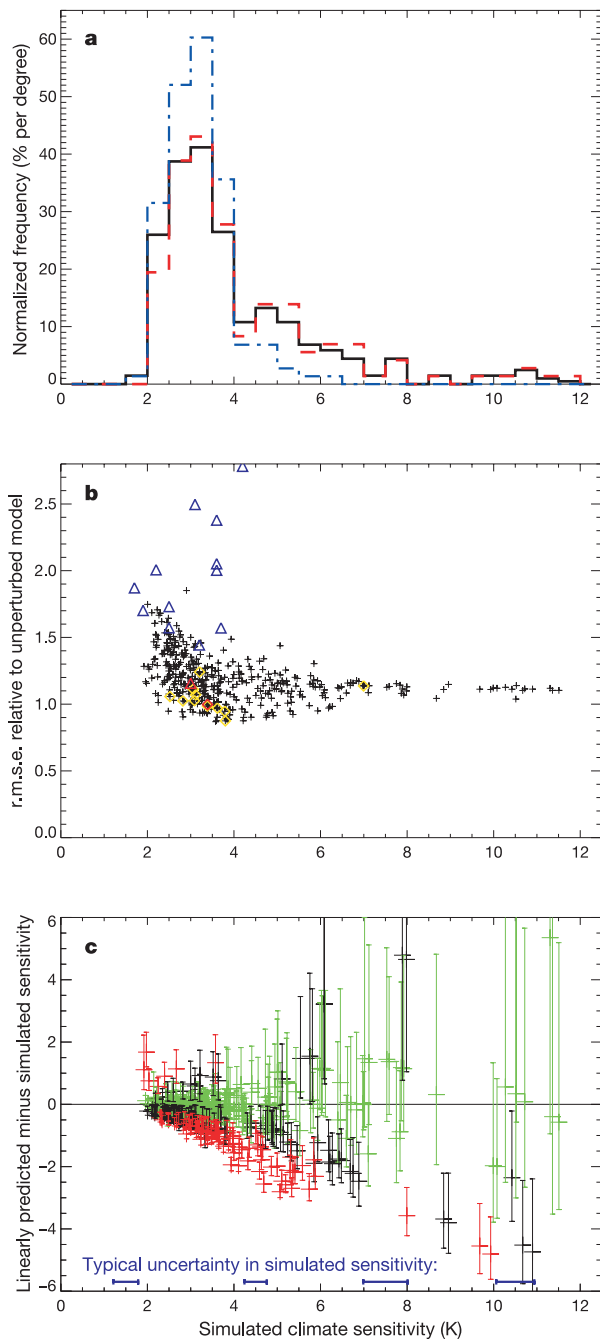


Figure 2 The response to parameter perturbations. **a**, The frequency distribution of simulated climate sensitivity using all model versions (black), all model versions except those with perturbations to the cloud-to-rain conversion threshold (red), and all model versions except those with perturbations to the entrainment coefficient (blue). **b**, Variations in the relative r.m.s.e. of model versions. The unperturbed model is shown by the red diamond. Model versions with only a single parameter perturbed are highlighted by yellow diamonds. The triangles show the CMIP II models for which data are available; HadCM3 (having the same atmosphere as the unperturbed model but with a dynamic ocean) is shown in red and the others in blue. **c**, Linear prediction of climate sensitivity based on summing the change in λ for the relevant single-parameter-perturbation model versions, to estimate λ when multiple perturbations are combined. Error bars show the resulting uncertainty ($\pm$ one sigma) caused by the combination of a number of $\Delta\lambda$ values where each λ has an uncertainty deduced from the initial-condition ensembles having only a single parameter perturbed. Linear predictions within one sigma of the simulated value are shown in green, between one and two sigma in black, and above two sigma in red. Mean uncertainties in the simulated value (two-sigma range, inferred from the initial-condition ensembles) are shown at the bottom for four regions of sensitivity (0–3, 3–6, 6–9, 9–12).

explored have relatively little effect on this global variable. There are a number of possible reasons for this clustering: the relevant processes may in fact have only a limited impact on sensitivity, the parameter ranges used may be too small to influence substantially the response in this model, and/or multiple perturbations may have mutually compensating effects when averaged on global scales. Of course, many significant regional impacts are invisible in a global average.

The range of sensitivities across different versions of the same model is more than twice that found in the GCMs used in the IPCC Third Assessment Report¹⁴. The possibility of such high sensitivities has been reported by studies using observations to constrain this quantity^{9,11,24,25}, but this is the first time that GCMs have generated such behaviour. The shape of the distribution is determined by the parameters selected for perturbation and the perturbed values chosen, which were relatively arbitrary. Model developers provided plausible high and low values for each model parameter; however, we cannot interpret these as absolute upper and lower bounds because experts are known to underestimate uncertainty even in straightforward elicitation exercises where the import of the question is clear²⁶. In our case even the physical interpretation of many of these parameters is ambiguous²⁷. We can illustrate the importance of the parameter choices by subsampling the model versions. If all perturbations to one parameter (the cloud-to-rain conversion threshold) are omitted, the red histogram in Fig. 2a is obtained, with a slightly increased fraction (4.9%) of model versions >8 K. If perturbations to another parameter (the entrainment coefficient) are omitted, the blue histogram in Fig. 2a is obtained, with no model versions >8 K. (See Supplementary Information for further sensitivity analyses.)

Can either high-end or low-end sensitivities be rejected on the basis of the model-version control climates? Fig. 2b suggests not; it illustrates the relative ability of model versions to simulate observations using a global root-mean-squared error (r.m.s.e.) normalized by the errors in the unperturbed model (see Methods). For all model versions this relative r.m.s.e. is within (or below) the range of values for other state-of-the-art models, such as those used in the second Coupled Model Inter Comparison (CMIP II) project²⁸ (triangles). The five variables used for this comparison are each standard variables in model evaluation and inter-comparison exercises²⁹ (see Methods). This lack of an observational constraint, combined with the sensitivity of the results to the way in which parameters are perturbed, means that we cannot provide an objective probability density function for simulated climate sensitivity. Nevertheless, our results demonstrate the wide range of behaviour possible within a GCM and show that high sensitivities cannot yet be neglected as they were in the headline uncertainty ranges of the IPCC Third Assessment Report (for example, the 1.4–5.8 K range for 1990 to 2100 warming).¹⁴ Further, they tell us about the sensitivities of our models, allowing better-informed decisions on resource allocation both for observational studies and for model development.

Can we coherently predict the model's response to multiple parameter perturbations from a small number of simulations each of which perturbs only a single parameter⁹? The question is important because it bears on the applicability of linear optimization methods in the design and analysis of smaller ensembles. Figure 2c shows that assuming that changes in the climate feedback parameter¹⁴ λ combine linearly provides some insight, but fails in two important respects. First, combining uncertainties gives large fractional uncertainties for small predicted λ and hence large uncertainties for high sensitivities. This effect becomes more pronounced the greater the number of parameters perturbed. Second, this method systematically underestimates the simulated sensitivity, as shown in Fig. 2c, and consequently artificially reduces the implied likelihood of a high response. Furthermore, more than 20% of the linear predictions are more than two standard errors from the

simulated sensitivities. Thus, comprehensive multiple-perturbed-parameter ensembles appear to be necessary for robust probabilistic analyses.

Figure 3 shows the initial-condition ensemble-mean of the temperature and precipitation changes for years 8–15 after doubling CO₂ concentrations, for three model versions: (1) the unperturbed model; (2) a version with low sensitivity; and (3) a version with high sensitivity (see Supplementary Information for details of the control climates in these model versions). All three models show the familiar increased warming at high latitudes and the overall surface-temperature pattern scales with sensitivity. Even in the low-sensitivity model version the warming in certain regions is substantial, exceeding 3 K in Amazonia and 4 K in much of North

America. The precipitation field shows a greater variety of response. For instance, this particular low-sensitivity model version shows a region of substantially reduced precipitation east of the Mediterranean; something not evident in either the standard or high-sensitivity model versions shown. It is critical to note that model versions with similar sensitivities often also show differences in such regional details⁹. The use of a GCM-based grand ensemble allows the significance of such details to be ascertained.

Thanks to the participation and enthusiasm of tens of thousands of individuals world-wide we have been able to discover GCM versions with comparatively realistic control climates and with sensitivities covering a much wider range than has ever been seen before. These results are a critical step towards a better under-

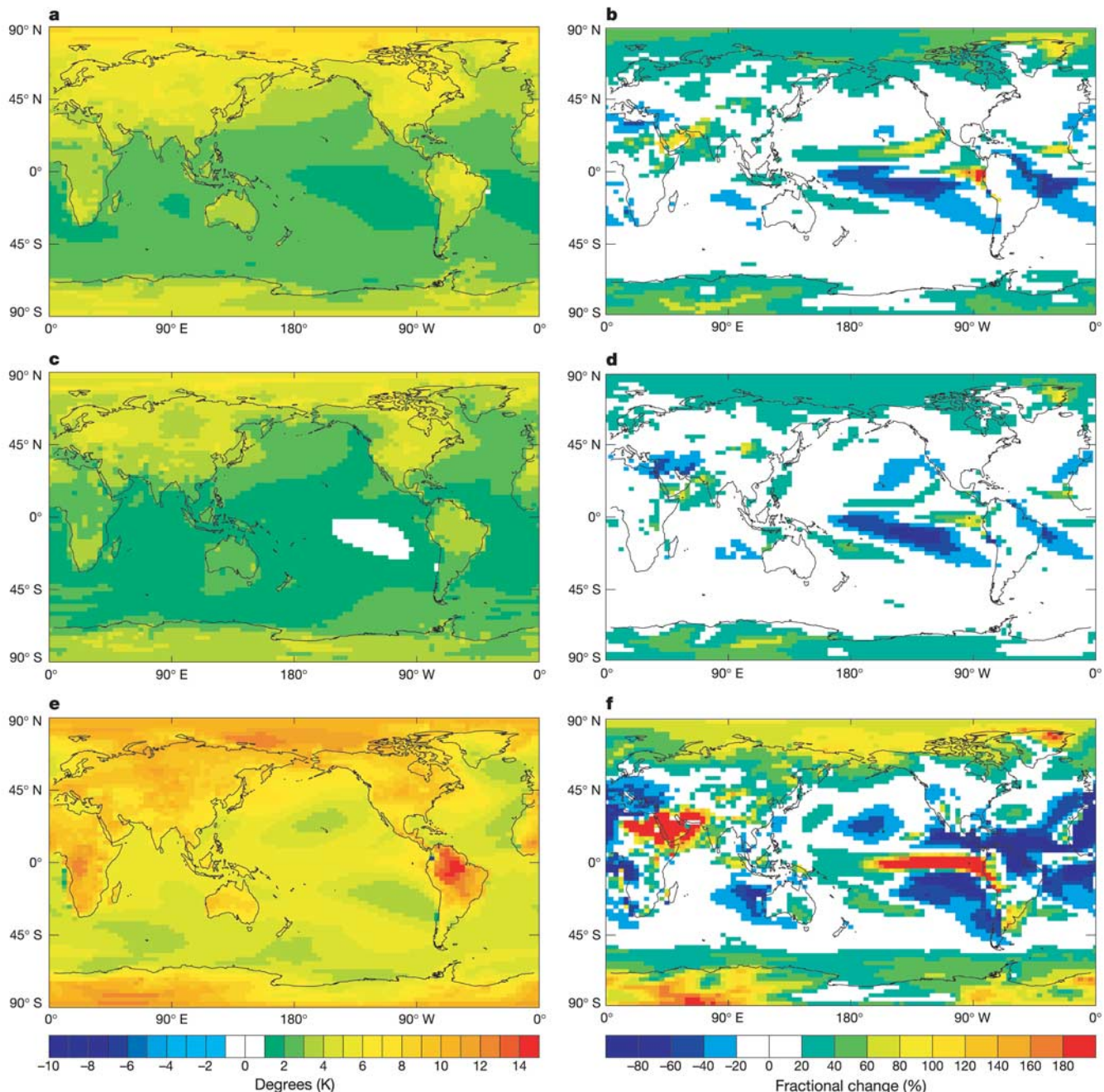


Figure 3 The temperature (left panels) and precipitation (right panels) anomaly fields in response to doubling the CO₂ concentrations. **a, b**, The unperturbed model (simulated climate sensitivity, 3.4 K). **c, d**, A model version with low simulated climate sensitivity

(2.5 K). **e, f**, A model version with high simulated climate sensitivity (10.5 K). These fields are the means of years eight to fifteen after the change of forcing is applied, averaged over initial-condition ensemble members; they are not the equilibrium response.

standing of the potential responses to increasing levels of greenhouse gases, regional and seasonal impacts, our models and internal variability. Future experiments will include a grand ensemble of transient simulations of the years 1950–2100 using a model with a fully dynamic ocean. □

Methods

Model simulations

Participants in the *climateprediction.net* experiment download an executable version of a full GCM. They are allocated a particular set of parameter perturbations and initial conditions enabling them to run one simulation: that is, one member of the grand ensemble. Their personal computer then carries out 45 years of simulation and returns results to the project's servers. Over 90,000 participants from more than 140 countries have registered to date. The model, based on HadSM3³⁰, is a climate resolution version of the Met Office Unified Model with the usual horizontal grid of 3.75° longitude × 2.5° latitude and 19 layers in the vertical. The ocean consists of a single thermodynamic layer with ocean heat transport prescribed using a heat-flux convergence field that varies with position and season but has no inter-annual variability. For each simulation the heat-flux convergence field is calculated in the calibration phase where sea surface temperatures (SSTs) are fixed; in subsequent phases the SSTs vary according to changes in the atmosphere–ocean heat flux. The initial-condition ensemble members have different starting conditions for the calibration and therefore allow for uncertainty in the heat-flux convergence fields used in the control and doubled-CO₂ phases.

Data quality

Most model simulations are unique members of the grand ensemble, each being a combination of perturbed model parameters and perturbed initial conditions. To evaluate the reliability of the experimental design a certain number of identical simulations are distributed; most give identical results. Where they do not, they are usually very similar, suggesting that a few computational bits were lost at some point and consequently they are essentially different members of the initial-condition ensemble. In these cases the mean of the simulations is taken.

There are a small number of simulations (1.6%) which show obvious flaws in the data: for example, sudden jumps of data values from of the order of 10² to of the order of 10⁸. These probably result from loss of information, for instance during a PC shut-down at a critical point in processing or a result of machine 'overclocking'. These are removed from this analysis. Finally, runs that show a drift in T_g greater than 0.02 Kyr⁻¹ in the last eight years of the control are judged to be unstable and are also removed from this analysis.

Perturbations

Perturbations are made to six parameters, chosen to affect the representation of clouds and precipitation: the threshold of relative humidity for cloud formation, the cloud-to-rain conversion threshold, the cloud-to-rain conversion rate, the ice fall speed, the cloud fraction at saturation and the convection entrainment rate coefficient. This is a subset of those explored by ref. 9. In each model version each parameter takes one of three values (the same values as those used by ref. 9); for cloud fraction at saturation only the standard and intermediate values are used. As *climateprediction.net* continues, the experiment is exploring 21 parameters covering a wider range of processes and values.

Climate sensitivity calculations

The simulated climate sensitivity is taken as the difference between the predicted equilibrium T_g in the doubled-CO₂ and control phases. The latter is simply the mean of the last eight years of that phase. The former is deduced by fitting the change in T_g relative to the start of the phase, to the exponential expression: $\Delta T_g(t) = \Delta T_{g(2\times CO_2)}(1 - \exp(-t/\tau))$, giving us a value of $T_{g(2\times CO_2)}$ that allows for uncertainty in the response timescale, τ . Even for high simulated climate sensitivities the uncertainty in this procedure is small (see Fig. 2c) and alternative methods give similar results. Because it is based on the first 15 years' response, the λ associated with this simulated climate sensitivity reflects the decadal timescale feedbacks in the system. Longer, centennial-timescale processes could affect the ultimate value of the equilibrium sensitivity and are best studied using models with dynamic oceans and cryospheres.

Relative root-mean-square error

Models are compared with gridded observations of annual mean temperature, sea level pressure, precipitation and atmosphere–ocean sensible and latent heat flux. The total error in variable j is defined simply as:

$$\varepsilon_{js}^2 = (\sum_i w_i (m_{is} - o_i)^2) \quad (1)$$

where m_{is} is the simulated value in grid-box i averaged over the last 8 yr of the control phase of simulation s , o_i is the observed value⁹ and w_i is an area weighting. Mean squared errors relative to the standard model are computed as:

$$\varepsilon_s^2 = (\sum_j \varepsilon_{js}^2 / \varepsilon_{ju}^2) / N \quad (2)$$

where N is the number of variables and ε_{ju}^2 is the mean ε_{js}^2 for the unperturbed model, and averaged across initial-condition ensembles. Normalizing errors in individual variables by the corresponding errors in the unperturbed model ensures that all variables are given equal weight. The relative r.m.s.e. is plotted in Fig. 2b. Note that because we do not have an explicit and adequate noise model (ε_{js}^2 does not account for correlations, for example), these 'scores' cannot be interpreted explicitly in terms of

likelihood, but nevertheless provide an indication of the relative merits of different model control climates.

For the CMIP II data the $(m_i - o_i)^2$ term is reduced by the variance of the mean to compensate for the greater variability found in models with dynamic oceans.

Received 4 November; accepted 20 December 2004; doi:10.1038/nature03301.

- Schneider, S. H. What is dangerous climate change? *Nature* **411**, 17–19 (2001).
- Reilly, J. *et al.* Uncertainty in climate change assessments. *Science* **293**, 430–433 (2001).
- Wigley, T. M. L. & Raper, S. L. Interpretation of high projections for global mean warming. *Science* **293**, 451–454 (2001).
- Allen, M. R. & Stainforth, D. A. Towards objective probabilistic climate forecasting. *Nature* **419**, 228 (2002).
- Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrological cycle. *Nature* **419**, 224–232 (2002).
- Palmer, T. N. Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**, 71–116 (2000).
- Smith, L. What might we learn from climate forecasts? *Proc. Natl Acad. Sci. USA* **99**, 2487–2492 (2002).
- Kennedy, M. C. & O'Hagan, A. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 425–450 (2001).
- Murphy, J. M. *et al.* Quantifying uncertainties in climate change from a large ensemble of general circulation model predictions. *Nature* **430**, 768–772 (2004).
- Knutti, R., Stocker, T. F., Joos, F. & Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).
- Forest, C. E. *et al.* Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117 (2002).
- Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. L. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **417**, 617–620 (2000).
- Stott, P. A. & Kettleborough, J. A. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726 (2002).
- Cubasch, U. *et al.* in *Climate Change 2001, The Science of Climate Change* (ed. Houghton, J. T.) Ch. 9 527–582 (Cambridge Univ. Press, Cambridge, 2001).
- Allen, M. R. Do-it-yourself climate prediction. *Nature* **401**, 627 (1999).
- Stainforth, D. *et al.* Distributed computing for public-interest climate modeling research. *Comput. Sci. Eng.* **4**, 82–89 (2002).
- Hansen, J. A. *et al.* Casino-21: Climate simulation of the 21st century. *World Res. Rev.* **13**, 187–189 (2001).
- Stainforth, D. *et al.* *Climateprediction.net*: Design principles for public-resource modeling. In *14th IASTED Int. Conf. on Parallel And Distributed Computing And Systems* (4–6 November 2002) 32–38 (ACTA Press, Calgary, 2002).
- Stainforth, D. *et al.* in *Environmental Online Communication* (ed. Scharl, A.) Ch. 12 (Springer, London, 2004).
- Giorgi, F. & Francisco, R. Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.* **27**, 1295–1298 (2000).
- Palmer, T. N. & Raisanen, J. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 512–517 (2002).
- Collins, M. & Allen, M. R. Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Clim.* **15**, 3104–3109 (2002).
- Pope, V. D., Gallani, M., Rowntree, P. R. & Stratton, R. A. The impact of new physical parameterisations in the Hadley Centre climate model – HadAM3. *Clim. Dyn.* **16**, 123–146 (2000).
- Gregory, J. M. *et al.* An observationally-based estimate of the climate sensitivity. *J. Clim.* **15**, 3117–3121 (2002).
- Andronova, N. G. & Schlesinger, M. E. Objective estimation of the probability density function for climate sensitivity. *J. Geophys. Res.* **106**, 22605–22612 (2001).
- Morgan, M. G. & Keith, D. W. Climate-change — subjective judgments by climate experts. *Environ. Sci. Technol.* **29**, 468–476 (1995).
- Smith, L. A. in *Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems* (ed. Mees, A. I.) Ch. 2 (Birkhauser, Boston, 2000).
- Covey, C. *et al.* An overview of results from the coupled model intercomparison project. *Glob. Planet. Change* **37**(1–2), 103–133 (2003).
- McAveney, B. J. *et al.* in *Climate Change 2001, The Science of Climate Change* (ed. Houghton, J. T.) Ch. 8, 471–524 (Cambridge Univ. Press, Cambridge, 2001).
- Williams, K. D., Senior, C. A. & Mitchell, J. F. B. Transient climate change in the Hadley centre models: the role of physical processes. *J. Clim.* **14**, 2659–2674 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank all participants in the 'climateprediction.net' experiment and the many individuals who have given their time to make the project a reality and a success. This work was supported by the Natural Environment Research Council's COAPEC, e-Science and fellowship programmes, the UK Department of Trade and Industry, the UK Department of the Environment, Food and Rural Affairs, and the US National Oceanic and Atmospheric Administration. We also thank Tessella Support Services plc, Research Systems Inc., Numerical Algorithms Group Ltd, Risk Management Solutions Inc. and the CMIP II modelling groups.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.A.S. (d.stainforth1@physics.ox.ac.uk).

Spectroscopic evidence for a lava fountain driven by previously accumulated magmatic gas

Patrick Allard^{1,2}, Mike Burton¹ & Filippo Muré¹

¹INGV, Piazza Roma 2, 95123 Catania, Italy

²Laboratoire Pierre Süe, CNRS-CEA, CE-Saclay, 91191 Gif sur Yvette, France

Lava fountains are spectacular continuous gas jets, propelling lava fragments to heights of several hundred metres, which occasionally occur during eruptions of low-viscosity magmas^{1–5}. Whether they are generated by the effervescent disruption of fast-rising bubbly melt^{2–5} or by the separate ascent of a bubble foam layer accumulated at depth^{6,7} still remains a matter of debate⁸. No field measurement has yet allowed firm discrimination between these two models. A key insight into the origin of lava fountains may be gained by measuring the chemical composition of the driving gas phase. This composition should differ markedly depending on whether the magma degassing occurs before or during eruption^{9,10}. Here we report the analysis of magmatic gas during a powerful (250–600 m high) lava fountain, measured with Fourier transform infrared spectroscopy^{11–14} on Mount Etna, Sicily. The abundances of volcanic gas species, determined from absorption spectra of lava radiation, reveal a fountain gas having higher CO₂/S and S/Cl ratios than other etnean emissions^{14–18}, and which cannot derive from syn-eruptive bulk degassing of Etna basalt^{19,20}. Instead, its composition suggests violent emptying of a gas bubble layer previously accumulated at about 1.5 km depth below the erupting crater.

Lava fountains have been principally described and studied on hotspot Hawaiian volcanoes^{1–5}, but are also observed on basaltic and alkaline volcanoes in other tectonic settings, such as Mount Etna, in Sicily. In January–June 2000, the Southeast summit crater (SEC, 3,250 m above sea level) of Etna produced an exceptional series of 64 periodic lava fountain episodes, with heights ranging from 80 to 900 m (ref. 21). These events, separated by quiescent intervals of 3 h to 10 days, had a reproducible pattern, and were associated with sharp peak increases of the seismic tremor²¹. The sixty-third eruptive episode, on 14 June 2000, was the best studied one²². After 13 h of precursory effusive then strombolian activity, it culminated with 40 min of fire fountaining that progressively increased in height from 250 to ~600 m, before ending abruptly.

During this event we could remotely measure the composition of the magmatic gas phase, by operating an open-path Fourier transform infrared (OP-FTIR) spectrometer (see Methods) at 990 m slanting distance from the south rim of SEC (Fig. 1a). OP-FTIR spectroscopy has recently been applied to analyse volcanic fumaroles¹¹ and plumes^{12–14}, but not yet a lava fountain. The fountain, several tens of metres wide, consisted of a red-orange core of incandescent gas and molten lava clots (~1,100 °C), surrounded by a darker envelope of finer pyroclasts (cinders, lapilli) and gas (Fig. 1b). The 7.5-mrad field of view of our telescope allowed us to target an ~8-m-wide area of this outer layer while collecting one FTIR spectrum every ~7 s. The 91 spectra collected during the fountain growth and climax (09:18 to 09:43 GMT) were pure absorption spectra, demonstrating that pyroclasts in the targeted area were always hotter than the coexisting volcanic gas. The mean temperature of these radiating clasts, determined by spectral fitting with a scaled Planck curve, varied between 510 and 630 °C, compared with 450 ± 25 °C for the gas, as inferred from the rotational band structure of the SO₂ absorption. Such moderate temperatures demonstrate extensive cooling of both solid particles and volcanic

gas in the outer part of a lava fountain, owing to radiative cooling and air entrainment^{2,5}.

The identity and molar path amount (in molecules per cm²) of each gas species along the beam path length were determined from spectral absorption lines of the infrared radiation emitted by pyroclasts. In each spectrum we detected eight gas species—namely H₂O, CO₂, SO₂, CH₄, N₂O, HCl, HF and CO (given in order of decreasing abundance). Whereas N₂O is purely atmospheric and SO₂, HCl and HF are purely volcanic, the other four species are present both in volcanic emissions and in air. Their amounts were thus corrected for the atmospheric amount of each species along the beam path length, derived from the γ -intercept of regression lines of scatter plots versus SO₂ and/or HCl (ref. 12). The gradients of regression lines provide the mean volcanic gas ratios presented below. The mean air concentrations inferred at ambient pressure (~680 hPa) and temperature (15 °C) of the measurement—7,700 parts per million by volume, p.p.m.v., H₂O (~30% relative

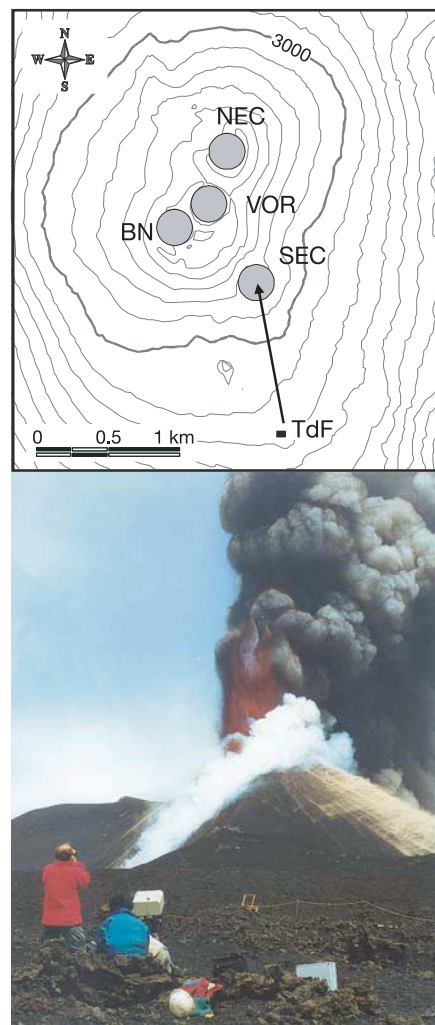


Figure 1 OP-FTIR measurement of the 14 June 2000 lava fountain on Etna. The sketch map shows the measuring site location, at 2,900 m elevation south of Southeast summit crater (SEC). TdF, Torre del Filosofo alpine shelter. VOR, BN and NEC denote Voragine, Bocca Nuova and Northeast summit craters, respectively. Beam path length is 990 m (arrow). The picture (35° angle view) shows the lava fountain at 09:28 GMT, when it was ~300 m high and lava had begun overflowing the south crater rim (whitish fumes in the foreground). The FTIR spectrometer is targeting an 8-m-wide area of the fountain about 70 m above the crater, in order to minimize interference from the whitish fumes.

humidity), 364 p.p.m.v. CO₂, 1.9 p.p.m.v. CH₄ and 0.12 p.p.m.v. CO—agree well with concentrations measured on top of Etna (H₂O, CO₂)¹⁸ and in Sicily (CO, CH₄)²³. Once air background is removed, most spectra contain pure volcanic CO₂ and CO, but no detectable methane, in agreement with the very low abundance of that species in magmatic gas¹⁰. The corrected concentrations for H₂O are less clear-cut: natural atmospheric variability increases their apparent error to ~20%, compared with 4–10% for other species; moreover, they can include unresolved and variable additions from air moisture entrained within the fountain, and possibly also ground water sucked into the erupting conduit. Therefore, we can only provide upper limits for the H₂O content of the magmatic gas.

Figure 2 shows the evolution of the gas concentrations and ratios together with seismic tremor amplitude, a good proxy for the eruption intensity²². Restricted chemical variations were initially observed during fountain growth (09:23:06 to 09:26:42 GMT; Fig. 2a). The driving gas contains on average 96 mol% H₂O, 3.7 mol% CO₂, 0.4 mol% SO₂, 0.037 mol% HCl, 0.02 mol% HF and 0.004 mol% CO. It has high mean CO₂/S (~10) and S/Cl (~10) ratios, and a low Cl/F ratio (~2). During increasingly powerful fountaining (09:28:48 to 09:40:47 GMT), the gas species displayed larger oscillations (Fig. 2a). These reflect the accelerating eruption dynamics, but also variable spectroscopic interference from whitish fumes rising from lava that started to overflow the south rim of SEC from 09:26 GMT (Fig. 1). This interference resulted in

occasional sharp increases in the optical thickness of the volcanic gas layer, but is easily identified by its distinct composition (Fig. 3). In this interval, the uncontaminated fountain gas had an average composition of 92 mol% H₂O, 7.3 mol% CO₂, 1.0 mol% SO₂, 0.10 mol% HCl, 0.07 mol% HF and 0.004 mol% CO. It maintained a similarly high S/Cl ratio (10), but had somewhat lower CO₂/S (7.3) and Cl/F (1.3) ratios than during the previous stage. We assessed its equilibrium temperature from the CO/CO₂ ratio and standard thermodynamic data for the reaction 2CO+O₂=2CO₂, assuming gas–melt equilibrium under redox conditions ($\log f_{O_2} \approx NNO+0.3$, where NNO is the nickel–nickel oxide oxygen buffer) typical for Etna basalt^{15,24}. This leads to: $T = -2,330 / [\log(CO/CO_2) + 0.357] - 273$, where T is in °C. The computed mean values of 640 ± 30 °C and 530 ± 70 °C for the two periods described above are higher than the physical gas temperature (450 ± 25 °C) inferred from SO₂ rotational structure, and close to the estimated Planck temperature of radiating pyroclasts (Fig. 2b), indicating efficient compositional quenching of the gas upon cooling.

Between 09:45 and 09:57 GMT we obtained a further 90 spectra of post-fountaining magma degassing at the south crater rim. In these emissions, only volcanic SO₂, HCl and HF were detected. The observed gradual decrease of S/Cl (from ~3.5 to 1) and Cl/F (from 1.3 to 0.9) ratios, along with declining activity (Fig. 2b), demonstrates a preferential exhaustion of SO₂ relative to HCl and HF during degassing of the cooling magma, as also observed

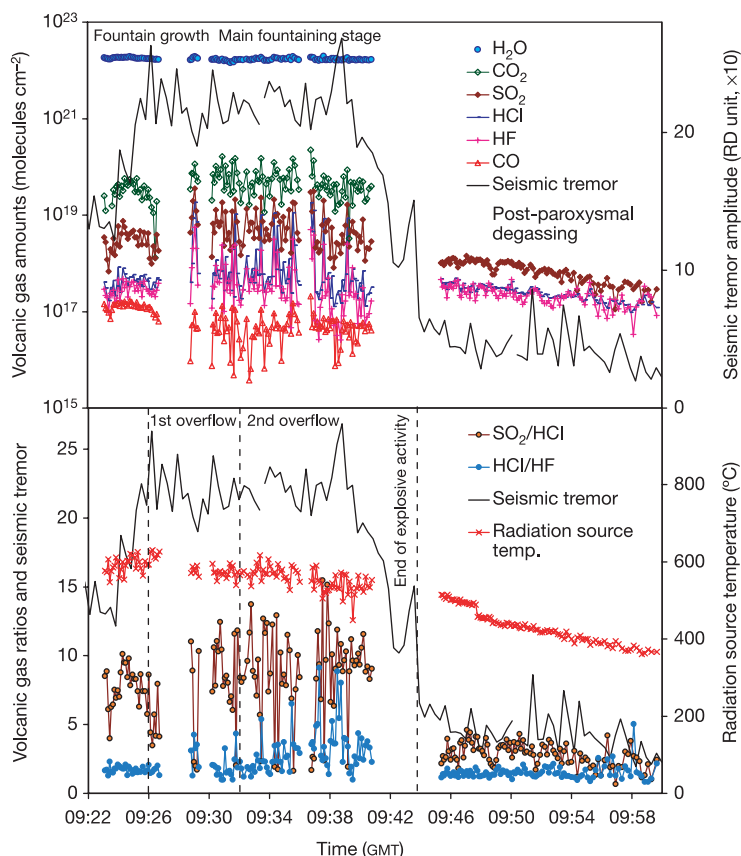


Figure 2 Temporal evolution of volcanic gas concentrations and ratios during and after the 14 June 2000 lava fountain. Top, time series of measured path amounts of volcanic gases (left-hand log scale), together with the volcanic tremor amplitude^{21,22} (right-hand scale, RD $\times 10$: here RD indicates reduced displacement, in cm²) as an indicator of the eruption intensity. H₂O, CO₂ and CO amounts are corrected for atmospheric background (see text). Bottom, evolution of SO₂/HCl and HCl/HF ratios (left-hand scale) during and

after the fountaining, and evolution of the radiating source temperature retrieved from FTIR spectra (right-hand scale). Smooth decrease of that temperature during the main fountaining stage reflects the influence of progressive widening of the falling (cooler) tephra area along with greater height of the jets. Left and centre dashed lines indicate the onset of bursts of lava overflowing the upper south crater rim; right dashed line indicates end of explosive activity.

from etnean lava flows^{14,15}. SO_2 was no longer detected in residual, HF-dominated emanations from the flowing lava, measured between 10:25 and 11:03 GMT.

Our data thus provide a detailed record of the chemical evolution of magmatic gas during and after a powerful lava fountain on a basaltic volcano. Before our study, one single spectroscopic datum had been obtained²⁵ in 1968 for H_2O , CO_2 and SO_2 during 10-m-high lava spattering on Kilauea volcano, in Hawaii. Compared with other etnean emissions^{14–18}, the fountain gas analysed here shows remarkable compositional features, which allow us to elucidate its genesis by reference to data for dissolved volatiles in Etna magma^{19,20,24,26}. Its high H_2O content and low $\text{CO}_2/\text{H}_2\text{O}$ ratio (0.04–0.08) are not particularly unusual¹⁵, as etnean gases arise from exceptionally water-rich basalts (3.4 wt%)¹⁹. However, the significance of these two parameters is limited by the above-mentioned, quite large, uncertainties on H_2O . In contrast, the CO_2/S and S/Cl ratios of the fountain are highly discriminating. Both are 2 to 4 times higher than the representative ratios for time-averaged Etna plume emissions (~ 3 and 2–4, respectively^{14,15,18}; Fig. 3) and no such S/Cl ratio has previously been measured in etnean eruptive gas¹⁵. Moreover, these two ratios are sensitive tracers of Etna basalt decompression and degassing^{15,19,20}. During magma ascent, carbon dioxide exsolves much earlier than sulphur, and sulphur earlier and more extensively than chlorine, this evolution being best documented for S and Cl^{19,20}. The coexisting bulk gas phase thus evolves from very high CO_2/S and S/Cl ratios at high pressure down to low (molar) ratios of ~ 3 at atmospheric pressure^{14,18–20}. Therefore, a lava fountain driven by syn-eruptive bulk magma degassing is expected to display CO_2/S and S/Cl ratios of

~ 3 , whereas previous gas–melt separation at depth should result in higher ratios.

Clearly, our results do not support a bulk degassing process and instead indicate gas–melt separation. We cannot exclude the possibility that degassing of fast-rising basalt could favour high S/Cl ratios owing to a lower diffusivity of chlorine relative to sulphur. However, the constant S/Cl ratio of the fountain gas, despite large variations in eruptive velocities²², as well as the lack of anomalous residual Cl content in the erupted tephra (P.A. *et al.*, manuscript in preparation), point to a minor influence of such late-stage kinetic effects. Moreover, earlier exsolved carbon dioxide and sulphur should be much less sensitive to such effects than more soluble HCl. Consequently, the presence of both high CO_2/S and S/Cl ratios in the fountain (Fig. 3) strongly argues in favour of a prevalent control of the gas composition by gas–melt separation at depth. The measured S/Cl ratio of ~ 10 corresponds to bulk equilibrium with Etna basalt under pressures of the order of 30–40 MPa (ref. 20 and N. Spilliaert *et al.*, manuscript in preparation), implying that the erupted gas phase preserved a composition established at about 1.5 km lithostatic depth below SEC (for a mean rock density of $2,600 \text{ kg m}^{-3}$).

Hence, we conclude that the measured lava fountain was most probably driven by the separate ascent of a gas layer previously accumulated at moderate depth within the volcanic pile (3,300 m high). This conclusion is consistent with independent evidence that the eruption was supplied by a shallow magma body undergoing rapid crystallization, on top of which repeated growth and collapse of a bubble foam may have triggered periodic lava fountaining (P.A. *et al.*, manuscript in preparation). This interpretation is reminiscent of the bubble foam model⁷ for periodic lava fountaining on Kilauea volcano. However, not all lava fountains are necessarily powered by this mechanism⁸. Here we demonstrate that remote gas sensing using OP-FTIR spectroscopy opens new ways to better understanding of lava fountain dynamics. □

Methods

Our Bruker model OPAG-22 OP-FTIR spectrometer has been operated since May 2000 for routine solar occultation measurements of Etna's summit plume composition²⁷. It has a ZnSe beamsplitter, 0.5 cm^{-1} spectral resolution and 1.78 cm optical path difference, and is equipped with an internal black-body calibration system. The detector used was a liquid-nitrogen-cooled InSb photovoltaic semiconductor with effective sensitivity between $1,500$ and $5,000 \text{ cm}^{-1}$. Spectra were calculated from the Fourier transform of double-sided interferograms using Norton-Beer medium apodization and power spectrum phase correction. Gas concentrations were determined from spectra using a nonlinear least-squares fitting program based on the optimal estimation algorithm²⁸ and infrared absorption parameters from the HITRAN 96 spectral database²⁹. We considered a two layer model of atmospheric (15°C) and volcanic gases (80°C for lava fumes, 450°C for fountain gas), calculated from the RFM radiative transfer code (<http://www-atm.atm.ox.ac.uk/RFM>). The volcanic gas temperature was retrieved by analysing the rotational structure of the SO_2 absorption. The uncertainty due to measurement and retrieval on reported gas amounts ranges from 4% to 10%.

Received 11 July; accepted 29 November 2004; doi:10.1038/nature03246.

- Swanson, D. A., Duffield, W. A., Jackson, D. B. & Peterson, D. B. Chronological narrative of the 1969–1971 Mauna Ulu eruption of Kilauea volcano, Hawaii. *US Geol. Surv. Prof. Pap.* **1056**, 1–59 (1979).
- Wilson, L. & Head, J. W. Ascent and eruption of basaltic magma on the Earth and Moon. *J. Geophys. Res.* **86**, 2971–3001 (1981).
- Head, J. W. & Wilson, L. Lava fountain heights at Pu'u'O'o, Kilauea, Hawaii: indicators of amount and variations of exsolved magma volatiles. *J. Geophys. Res.* **92**, 13715–13719 (1987).
- Sparks, R. S. J. The dynamics of bubble formation and growth in magmas. *J. Volcanol. Geotherm. Res.* **3**, 1–37 (1978).
- Parfitt, E. A., Wilson, L. & Neal, C. A. Factors influencing the height of Hawaiian lava fountains: implications for the use of fountain height as an indicator of magma gas content. *Bull. Volcanol.* **57**, 440–450 (1995).
- Jaupart, C. & Vergnolle, S. Laboratory models of Hawaiian and Strombolian eruptions. *Nature* **331**, 58–60 (1988).
- Vergnolle, S. & Jaupart, C. Dynamics of degassing at Kilauea volcano, Hawaii. *J. Geophys. Res.* **95**, 2793–2809 (1990).
- Parfitt, E. A. A discussion of the mechanisms of explosive basaltic eruptions. *J. Volcanol. Geotherm. Res.* **134**, 131–144 (2004).
- Gerlach, T. M. Exsolution of H_2O , CO_2 , and S during eruptive episodes at Kilauea Volcano, Hawaii. *J. Geophys. Res.* **91**, 12177–12185 (1986).

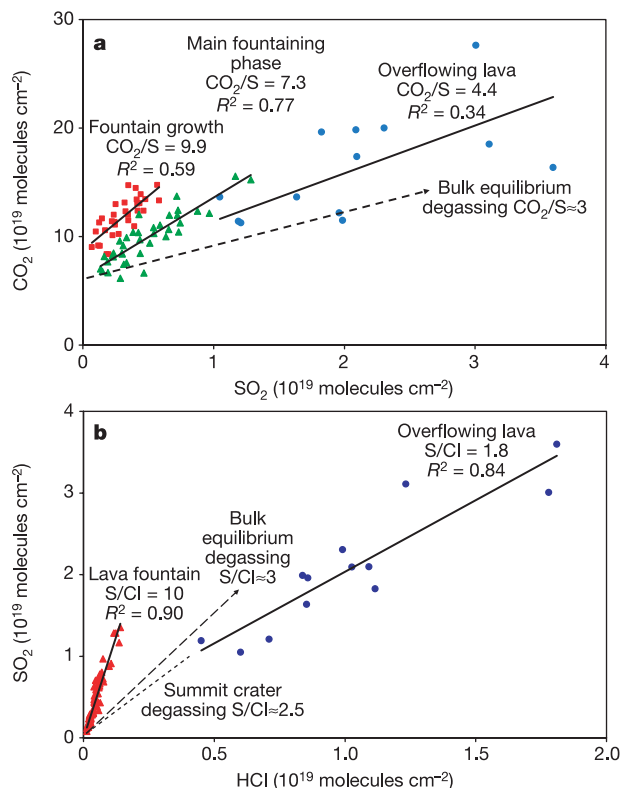


Figure 3 Scatter plots of gas amounts measured during lava fountaining. **a**, CO_2 versus SO_2 , and **b**, SO_2 versus HCl , showing the distinct mean volcanic ratios in fountain gas and whitish fumes from overflowing lava. Average S/Cl and CO_2/S ratios for bulk degassing of Etna basalt^{14,18,19} and S/Cl ratio measured in May–June 2000 summit plume emissions²⁷ are shown for comparison.

10. Giggenbach, W. F. In *Monitoring and Mitigation of Volcano Hazards* (eds Scarpa, R. & Tilling, R.) (Springer, Berlin, 1996).
11. Mori, T. *et al.* Remote detection of fumarolic gas chemistry at Vulcano, Italy, using an FT-IR spectral radiometer. *Earth Planet. Sci. Lett.* **134**, 219–224 (1995).
12. Francis, P. W., Burton, M. & Oppenheimer, C. Remote measurements of volcanic gas compositions by solar FTIR spectroscopy. *Nature* **396**, 567–570 (1998).
13. Burton, M., Oppenheimer, C., Horrocks, L. A. & Francis, P. W. Field measurement of CO₂ and H₂O emissions from Masaya Volcano, Nicaragua, by Fourier transform spectrometry. *Geology* **28**, 915–918 (2000).
14. Burton, M., Allard, P., Muré, F. & Oppenheimer, C. In *Volcanic Degassing* (eds Oppenheimer, C., Pyle, D. & Barclay, J.) 281–293 (Special Publication 2213, Geological Society, London, 2003).
15. Allard, P. *Géochimie Isotopique et Origine de l'eau, du Carbone et du Soufre dans les Gaz Volcaniques: Zones de Rift, Marges Continentales et Arcs Insulaires*. Thesis, Paris 7 Univ. (1986).
16. Allard, P. *et al.* Eruptive and diffuse emissions of carbon dioxide from Etna volcano. *Nature* **351**, 387–391 (1991).
17. Pennisi, M. & Le Cloarec, M. F. Variations of Cl, F, and S in Mount Etna's plume, Italy, between 1992 and 1995. *J. Geophys. Res.* **B 103**, 5061–5070 (1998).
18. Allard, P. *Monitoring Volcanic Risk by Remote Sensing Techniques: Airborne Plume Validation* (Final Rep. MVRSS RTD Contract, ENV4960288, CEE-DGXII, Bruxelles, 1999).
19. Métrich, N., Allard, P., Spilliaert, N., Andronico, D. & Burton, M. 2001 flank eruption of the alkali- and volatile-rich primitive melt responsible for Mount Etna's evolution in the last three decades. *Earth Planet. Sci. Lett.* **228**, 1–17 (2004).
20. Spilliaert, N., Allard, P., Métrich, N. & Sobolev, A. A new step in modelling magma degassing processes and eruptive dynamics at Mount Etna: the 2002–2003 eruption. *Geophys. Res. Abstr.* **6**, 04189 (2004).
21. Alparone, S., Andronico, D., Lodato, L. & Sgroi, T. Relationship between tremor and volcanic activity during the Southeast Crater eruption on Mount Etna in early 2000. *J. Geophys. Res.* **108**, 2241, doi:10.1029/2002JB001866 (2003).
22. Allard, P. *et al.* Source process of cyclic fire fountaining at Mt. Etna in 2000: a multidisciplinary study of the June 14 (63rd) event. *Geophys. Res. Abstr.* **5**, 13079 (2003).
23. World Data Centre for Greenhouse Gases, Japan Meteorological Agency. <http://gaw.kishou.go.jp/wdogg.html>.
24. Métrich, N. & Clocchiatti, R. Sulfur abundance and its speciation in oxidized alkaline melts. *Geochim. Cosmochim. Acta* **60**, 4151–4160 (1996).
25. Naughton, J. J., Derby, J. V. & Glover, R. B. Infrared measurements on volcanic gas and fume: Kilauea eruption, 1968. *J. Geophys. Res.* **74**, 3273–3277 (1969).
26. Métrich, N., Clocchiatti, R., Mosbah, M. & Chaussidon, M. The 1989–1990 activity of Etna magma mingling and ascent of H₂O-Cl-S-rich basaltic magma. Evidence from melt inclusions. *J. Volcanol. Geotherm. Res.* **59**, 131–144 (1993).
27. Allard, P., Burton, M., Brusca, L. & Muré, F. *Geochemical Monitoring of Sicilian Volcanoes* (Second Rep., July–December 2000, Sistema Poseidon, Nicolosi, Italy, 2000).
28. Rodgers, C. D. Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys. Space Phys.* **14**, 609–624 (1976).
29. Rothman, L. S. *et al.* The HITRAN molecular spectroscopic database and HAWKS (HITRAN Atmospheric Workstation): 1996 edition. *J. Quant. Spectrosc. Radiat. Transfer* **60**, 665–710 (1998).

Acknowledgements We thank F. Barberi and L. Villari for their initial support of the Etna FTIR monitoring project within the Poseidon System (Italian Civil Defence and Sicily Region), and A. Bonaccorso, E. Boschi and S. Calvari for their continued support within INGV. We acknowledge early technical assistance from the Bruker company (Germany) and discussions with N. Métrich (LPS) and our colleagues in INGV-Catania.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.A. (patrick.allard@cea.fr).

Long-term relationships between ecological stability and biodiversity in Phanerozoic reefs

Wolfgang Kiessling

Museum of Natural History, Humboldt University Berlin, Invalidenstrasse 43, D-10115 Berlin, Germany

High biodiversity has been shown to enhance ecological stability on small spatial scales and over intervals of weeks to decades^{1–4}. It remains unclear, however, whether this diversity–stability relationship can be scaled up to regional scales, or to longer timescales⁵. Without empirical validation at larger scales, the implications of the diversity–stability relationship for both ecology and long-term conservation strategies cannot readily

be resolved. Here I show that in biogenic reefs, ecological stability is related to taxonomic diversity on million-year timescales. The higher the mean reef diversity in a particular time interval, the smaller the change in skeletal density, style of reef building and biotic reef types in the subsequent time interval. Because the relationships apply to a wide spectrum of disturbance regimes and reef types, these results support the hypothesis that species richness itself promotes ecological stability³. Carbonate production by reefs, while closely correlated with reef diversity without temporal lag, is not stabilized by reef diversity over these long timescales. This suggests that ecological stability and productivity may be decoupled in natural ecosystems.

Reefs, broadly defined as laterally confined structures built by the growth and/or metabolic activity of sessile benthic organisms in an aquatic environment, are an excellent tool for tracking long-term ecological changes, because they have an outstanding fossil record reaching back at least two billion years and because they yield a number of attributes, other than taxonomic richness, that permit the identification of ecological traits. Reefs changed substantially in abundance, composition, palaeogeographic distribution, geometric attributes and biodiversity during the Phanerozoic eon (the past 542 million years, Myr), but these changes are rarely linearly correlated with global environmental changes—at least as inferred from the geological record^{6,7}. Biological controls seem either to buffer the reef system (the combined traits of all reefs recorded within a time interval) or to amplify responses to global change on timescales of millions of years^{7,8}. Using a database of more than 3,300 Phaner-

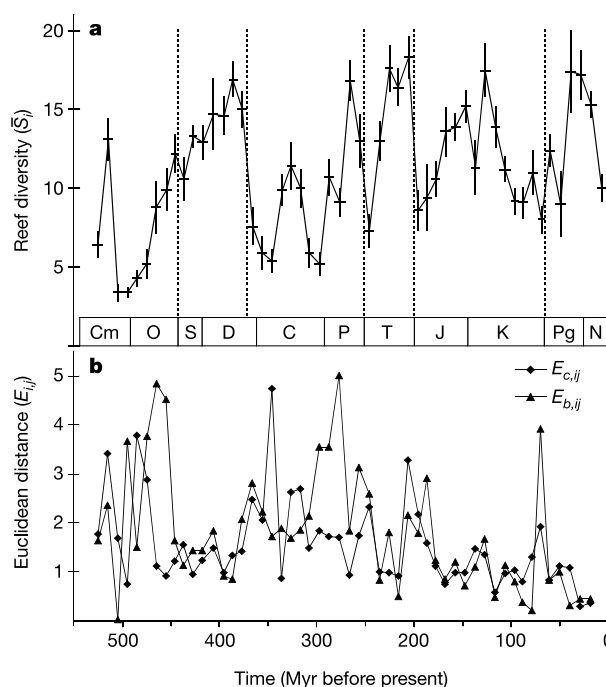


Figure 1 Time series of mean reef diversity and two measures of ecological change resolved to 10-Myr intervals. **a**, Reef diversity expressed as the mean species richness (number of species) of reef-builders within individual reef complexes per time interval. Vertical bars indicate one standard error of the mean in each direction. Horizontal bars demarcate the spans of time intervals. Vertical dashed lines mark the big five mass extinction events of the Phanerozoic. Cm, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene; N, Neogene. **b**, Ecological changes between consecutive time intervals ($i-j$, plotted in j) expressed as the euclidean distances of constructional styles ($E_{c,j}$) and biotic reef types ($E_{b,j}$).

ozoic reef complexes, I show that on these long timescales, reef diversity itself forms an important control of ecological stability.

Relationships between reef diversity and ecological stability were tested through correlations between reef diversity in one time interval (i) and the change of ecological traits between this initial interval and the succeeding interval (j). Reef diversity ($\bar{S}_i$) was measured by the average number of reef-building species within individual reefs per time interval (Fig. 1a). Ecological change (D_{ij}) was assessed by euclidean distances (E_{ij} , Fig. 1b) and Mahalanobis distances (M_{ij}) among variables divided into four sets (see Methods) that reflect: (1) changes in the density of skeletal reef builders (d), relative to matrix and cement, $E_{d,ij}$ and $M_{d,ij}$; (2) changes in reef architecture (a), $E_{a,ij}$; (3) changes in reef construction styles (c), a combination of (1) and (2), $E_{c,ij}$ and $M_{c,ij}$; and (4) changes in biotic reef types (b), $E_{b,ij}$. Changes in carbonate production (p), $D_{p,ij}$, were measured by the change in one variable: the volume of preserved reef carbonate per million years (ref. 9). Analyses of the relationship between reef diversity and ecological changes were performed on three time series at differing sample resolutions: (1) traditional stages or epochs (73 intervals), largely defined by global changes of fossil assemblages; (2) regularly spaced 10-Myr intervals (53 intervals); and (3) supersequences (32 intervals), defined by nearly global breaks in the sedimentary record¹⁰. Because reef diversity and several measures of ecological change are serially correlated, all time series were de-trended to $\Delta\bar{S}_i$ and ΔD_{ij} before correlation tests (see Methods). Owing to space constrictions, only 10-Myr time series are plotted. A full set of graphs is available as Supplementary Information.

Initial tests of the relationship between $\Delta\bar{S}_i$ and ΔD_{ij} rarely

yielded significant linear relationships. Except for $\Delta D_{p,ij}$, all plots of $\Delta\bar{S}_i$ versus ΔD_{ij} show a negative slope—implying that greater reef diversity tends to reduce ecological change—but several outliers prevent conclusive statements (Fig. 2). Exceptions are $\Delta\bar{S}_i$ versus $\Delta E_{c,ij}$, which, applying Spearman's Rho (r_s), is initially significant at all three temporal resolutions. Outliers, detected by least-trimmed square regressions and visual inspection of $\Delta\bar{S}_i$ – ΔD_{ij} plots, vary among analyses and temporal resolutions but are usually associated with major mass extinction events, or lie in the Cambrian period. Uncertainties of estimates of both reef diversity and ecological changes in the predominantly microbial reef system of the Cambrian, and the high volatility of diversity dynamics in this period^{11,12}, justify an exclusion from further analyses. The outlying mass extinctions are characterized by large ecological changes following times of high reef diversity. The ability of mass extinction events to cause disruption to regular evolutionary patterns has long been noted¹³. However, it is not reasonable to exclude all mass extinctions from the correlation analyses. Although mass extinctions may break the persistence stability (ability to withstand perturbations) on shorter timescales, resilience stability (potential for recovery from perturbations) on longer timescales is probably responsible for the observation that even the strongest mass extinctions are not recognized as outliers in all time series. The Permian–Triassic mass extinction provides an instructive example: at the stage/epoch level, great ecological distances are noted between all stages from the Late Permian to the Middle Triassic (Anisian), owing to a complete but transient destruction of metazoan reefs. At the supersequence level, however, great ecological similarities have been noted previously between Middle–Late Permian and Middle Triassic reefs¹⁴.

When the Cambrian period and outlying mass extinctions are excluded from the analyses, significant negative correlations between $\Delta\bar{S}_i$ and ΔD_{ij} are evident in all three time series and for nearly all measures of ecological change, except for $\Delta D_{p,ij}$ (Table 1). If reef diversity is higher than average in one time interval, lower

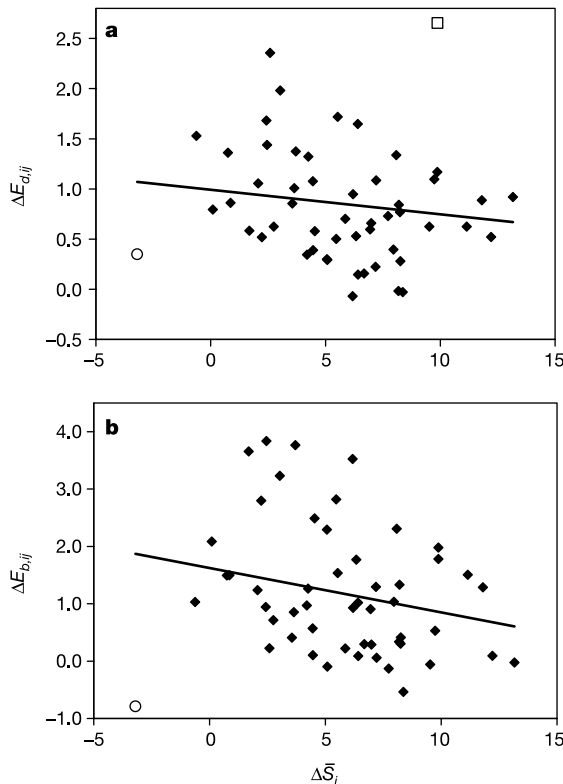


Figure 2 Examples of initial relationships between de-trended reef diversity ($\Delta\bar{S}_i$) and de-trended measures of ecological change in 10-Myr intervals. Marked outliers are: open circles, Middle–Late Cambrian boundary; open square, Triassic–Jurassic boundary (late Norian/Rhaetian–Hettangian/Sinemurian). **a**, Correlation with changes in skeletal framework density; **b**, correlation with changes in biotic reef types.

Table 1 Correlations between initial reef diversity and ecological changes

Ecological trait	r_s	P
Supersequences ($n = 27$)		
Framework density ($E_{d,ij}$)	−0.63	<0.001
Framework density ($M_{d,ij}$)	−0.53	0.005
Reef architecture ($E_{a,ij}$)	−0.70	<0.001
Constructional style ($E_{c,ij}$)	−0.76	<0.001
Constructional style ($M_{c,ij}$)	−0.56	0.002
Biotic reef type ($E_{b,ij}$)	−0.35	0.070
Carbonate production ($D_{p,ij}$)	0.20	0.314
Carbonate production ($ D_{p,ij} $)	0.43	0.025
10-Myr intervals ($n = 47$)		
Framework density ($E_{d,ij}$)	−0.36	0.012
Framework density ($M_{d,ij}$)	−0.34	0.020
Reef architecture ($E_{a,ij}$)	−0.50	<0.001
Constructional style ($E_{c,ij}$)	−0.59	<0.001
Constructional style ($M_{c,ij}$)	−0.32	0.027
Biotic reef type ($E_{b,ij}$)	−0.41	0.004
Carbonate production ($D_{p,ij}$)	0.11	0.461
Carbonate production ($ D_{p,ij} $)	0.27	0.066
Stages/epochs ($n = 65$)		
Framework density ($E_{d,ij}$)	−0.29	0.018
Framework density ($M_{d,ij}$)	−0.23	0.070
Reef architecture ($E_{a,ij}$)	−0.22	0.077
Constructional style ($E_{c,ij}$)	−0.34	0.006
Constructional style ($M_{c,ij}$)	−0.35	0.004
Biotic reef type ($E_{b,ij}$)	−0.29	0.020
Carbonate production ($D_{p,ij}$)	0.06	0.631
Carbonate production ($ D_{p,ij} $)	0.14	0.285

Spearman rank order correlations are based on normalized values and de-trended time series, and exclude the following intervals and boundaries. In the supersequences, I excluded the Cambrian (two intervals), the Frasnian–Famennian boundary (Givetian/Frasnian–Famennian/early Viséan) and the Triassic–Jurassic boundary (late Carnian/Rhaetian–Hettangian/early Aalenian). In the 10-Myr intervals, I excluded the Cambrian (four intervals) and the Triassic–Jurassic boundary (late Norian/Rhaetian–Hettangian/Sinemurian). In the stages/epochs, I excluded the Cambrian (five intervals), the Permian–Triassic boundary (Changhsingian–Scythian) and the Triassic–Jurassic boundary (Rhaetian–Hettangian/Sinemurian).

than average ecological changes occur in the subsequent time interval with respect to the density of skeletal organisms, reef architecture, reef construction style and biotic reef types. Although some individual variables defining ecological change are inevitably correlated with $\Delta\bar{S}_i$ (for example, $\bar{S}$ is significantly higher in intervals with many coral reefs than in intervals with a prevalence of microbial reefs), the combinations of variables used to define ecological distances are largely independent of reef diversity. Cross-correlations confirmed that significant correlations listed in Table 1 only occur at lag 1 (Fig. 3), that is, reef diversity in one time interval is related to ecological change in the subsequent interval, and not the converse.

The basic conclusion is that high reef diversity in one time interval predicts low ecological change in the subsequent time interval, whereas low-diversity reefs were more prone to ecological changes. The variance explained by $\Delta\bar{S}_i$ for individual measures of ΔD_{ij} is low enough (5%–58%) to leave space for environmental influences to control ecological changes, but, with the exception of major mass extinctions, environmental perturbations rarely forced substantial changes of large-scale ecological attributes of a high-diversity reef system. These results provide strong support for the existence of diversity–stability relationships on very long timescales, and also help to pinpoint their underlying causes. Because the diversity–stability relationship is evident through a wide range of reef types growing under variable physicochemical and biological disturbance regimes⁸, we may conclude that species richness itself had a stabilizing effect. This supports the diversity insurance hypothesis^{3,15}, which states that a high-diversity system is: (1) more likely to have species able to tolerate environmental perturbations; and (2) more likely to have functional redundancy so that ecologically important species can be readily replaced after a perturbation. The alternatives, that individual species traits and ecological history are more crucial to diversity–stability relationships^{16,17}, are unlikely for these long timescales.

The second conclusion is that reef diversity has little effect on long-term changes of carbonate production (Fig. 3). $\Delta\bar{S}_i$ is strongly correlated with concurrent carbonate production (Δp_{ij}) at all three sample resolutions ($r_s = 0.39$ – 0.52 , $P < 0.001$ – 0.004). This suggests that either reef carbonate production and reef diversity are driven by related controls, or that diversity itself promotes productivity. The latter view is known as the diversity–productivity hypothesis, which is supported by theoretical models and grassland experiments^{18,19}. However, the future development of reef carbonate production is apparently not predictable from reef diversity. There is even a tendency for absolute values of $\Delta D_{p,ij}$ to increase, after intervals with high $\Delta\bar{S}$ (Table 1). This is consistent with the results of

small-scale experiments where species-poor systems were shown to be more stable against perturbations in terms of biomass, although diversity and biomass were initially positively correlated²⁰. Here, the contradiction can be partly explained by the recognition of reefs as multi-stable ecosystems, able to achieve high rates of carbonate production with different consortia of communities²¹. Turning this argument around could mean that substantial fluctuations in carbonate production can occur in ecologically stable reefs; an interpretation supported by the record of coral reefs from the Pleistocene epoch²².

Because my results largely agree with findings from smaller-scale studies, it seems that diversity–stability relationships scale up from short observation periods and laboratory scales to millions of years and global scales. Although the validity of diversity–stability relationships at intermediate timescales—within the range of long-term ecosystem management—remains to be demonstrated, these analyses emphasize the urgency with which current threats to coral reef diversity and function^{23,24} need to be addressed. Although coral reefs are relatively diverse at present, this high biodiversity is not likely to protect them against productivity declines, because standing diversity appears to be uncoupled from productivity changes in these systems. Furthermore, if coral reefs and their biodiversity continue to decline as current trends predict, reef behaviour may eventually become unpredictable and unmanageable. The fossil record shows that even in the worst cases, reef destruction does not last forever, but a re-emerging reef system will probably look very different from today's tropical paradises. □

Methods

Database

Data were extracted from PaleoReefs, a palaeogeographic database of pre-Pleistocene Phanerozoic reefs that encompasses quantitative data on biogenic and petrographic composition, (palaeo-)geography, environmental setting and geometric attributes²⁵. An individual reef in the database is represented by an isolated reef of more than 20 km in maximum diameter, a reef transect 20 km in length (if the reef is more than 20 km in maximum diameter), or data lumped from reefs spaced less than 20 km apart (if the reefs are of the same age and environment).

The relevant variables for analyses are: diversity (see subsection 'Assessment of reef diversity'); matrix (carbonate mud or sand) and spar (carbonate cement) content, relative to the concentration of skeletal reef builders (ordinal values); architecture of a reef based on its inferred syndepositional relief, ranging from true rigid reef with maximum relief to biostromes with no relief (ordinal values); the dominant group of fourteen groups that construct reefs (supra-order groups, except for *Tubiphytes*, a reef builder of disputed taxonomic affinity; nominal values); and length, width and thickness (ordinal and scale values).

The database is available online at <http://193.175.236.205/paleo> (id = paleo, password = reefs). Taxonomic data of the database are deposited in the Paleobiology Database (<http://paleodb.org/>). All analyses exclude non-tropical reef types.

Assessment of reef diversity

Each diversity value in the database represents the total number of species, S , contributing to the construction of a reef (reef builders). The species richness of the dominant modern reef builders (scleractinian corals) has been shown to correlate with the diversity of reef-dwelling organisms²⁶ and thus may serve as a proxy of total reef diversity. Metric diversity values could be extracted for 650 reefs. Broad diversity intervals (low, moderate, high) were defined to include reefs with incomplete taxonomic data but with estimates of total diversity in the published literature (1,736 values). Ordinal values were transformed into scale values to permit the calculation of means on the basis of the empirical relationship: low diversity = 3 species; moderate diversity = 12 species; high diversity = 30 species. A maximum of 45 reef-building species was allowed for individual reefs, to adjust for monographic effects.

Assessment of ecological change

For euclidean distances, relevant attributes were first aggregated globally per time interval (mean or percentage values for ordinal values, percentage values for nominal values). Individual sets of variables comprise: (1) mean density, d , of skeletal reef builders (mean matrix, mean spar); (2) reef architecture, a (percentages of true reefs, reef mounds and mud mounds); (3) reef construction styles, c (variables of 1 and 2 combined); and (4) biotic reef type, b , expressed as the percentage of reefs dominated by a particular group of reef builders (microbes, calcareous algae, corals, calcareous sponges, siliceous sponges, bivalves, bryozoans or *Tubiphytes*). An euclidean distance matrix of standardized (rescaled to unit variance, z-scores) sets of variables was computed and the distances between consecutive time intervals (E_{ij}) were tabulated.

To adjust for covariances in the variables, Mahalanobis distances between individual populations of reefs within time intervals (M_{ij}) were also computed for sets of ordinal

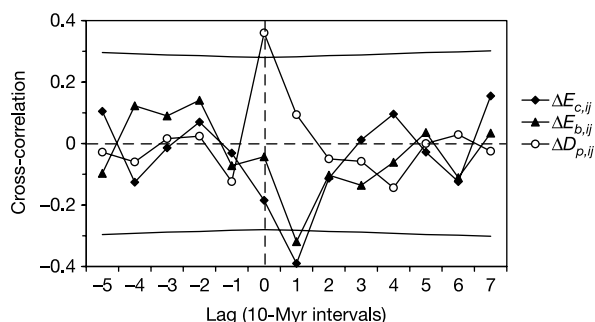


Figure 3 Cross-correlation between reef diversity and selected measures of ecological change in ten-Myr intervals. Cambrian omitted, but the outlier at the Triassic–Jurassic boundary is included to preserve a complete time series. Curved lines indicate the upper and lower 95% confidence bounds. Positive lags indicate ecological change lagging diversity.

variables (matrix content, spar content, reef architecture) after conversion to z-scores. Because this measure requires listwise deletion of data, considerably fewer data are available than for the computation of means and percentages.

In contrast to reef diversity and the other methods of ecological change, recorded reef volume is strongly dependent on the available rock record and exploration intensity, but it is unlikely that $D_{p,ij}$ is completely obscured by heterogeneities of the geological record⁹.

Time series

Stages or epochs with fewer than ten reefs were combined with the subsequent time interval. The 10-Myr intervals were slightly adjusted (± 2 Myr for individual intervals) to avoid lumping data across major extinction events (see Supplementary Information). Lag 1 autocorrelations were removed by calculating generalized differences of the time series data²⁷:

$$\Delta x = x_t - r_A(x_{t-1})$$

where x_t is the original value and r_A is the lag 1 autocorrelation. The first point in the time series was modified to:

$$\Delta x_1 = x_1 \sqrt{1 - r_A^2}$$

Received 24 April; accepted 25 October 2004; doi:10.1038/nature03152.

1. Tilman, D. & Downing, J. A. Biodiversity and stability in grasslands. *Nature* **367**, 363–365 (1994).
2. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
3. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
4. McCann, K. S. The diversity–stability debate. *Nature* **405**, 228–233 (2000).
5. Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
6. Webb, G. E. Was Phanerozoic reef history controlled by the distribution of non-enzymatically secreted reef carbonates (microbial carbonate and biologically induced cement)? *Sedimentology* **43**, 947–971 (1996).
7. Kiessling, W. in *Phanerozoic Reef Patterns* (eds Kiessling, W., Flügel, E. & Golonka, J.) 625–690 (SEPM Special Publication 72, Tulsa, 2002).
8. Wood, R. *Reef Evolution* (Oxford Univ. Press, Oxford, 1999).
9. Kiessling, W., Flügel, E. & Golonka, J. in *Carbonate Platform Systems: Components and Interactions* (eds Insalaco, E., Skelton, P. W. & Palmer, T. J.) 191–215 (Geological Society Special Publication 178, London, 2000).
10. Golonka, J. & Kiessling, W. in *Phanerozoic Reef Patterns* (eds Kiessling, W., Flügel, E. & Golonka, J.) 11–20 (SEPM Special Publication 72, Tulsa, 2002).
11. Kirchner, J. W. & Weil, A. Delayed biological recovery from extinctions throughout the fossil record. *Nature* **404**, 177–180 (2000).
12. Alroy, J. Are Sepkoski's evolutionary faunas dynamically coherent? *Evol. Ecol. Res.* **6**, 1–32 (2004).
13. Jablonski, D. Background and mass extinctions: the alternation of macroevolutionary regimes. *Science* **231**, 129–133 (1986).
14. Flügel, E. in *Phanerozoic Reef Patterns* (eds Kiessling, W., Flügel, E. & Golonka, J.) 391–463 (SEPM Special Publication 72, Tulsa, 2002).
15. Yachi, S. & Loreau, M. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 1463–1468 (1999).
16. Sankaran, M. & McNaughton, S. J. Determinants of biodiversity regulate compositional stability of communities. *Nature* **401**, 691–693 (1999).
17. Bellwood, D. R., Hoey, A. S. & Choat, J. H. Limited functional redundancy in high diversity systems: resilience and ecosystem function on coral reefs. *Ecol. Lett.* **6**, 281–285 (2003).
18. Tilman, D. The ecological consequences of changes in biodiversity: a search for general principles. *Ecology* **80**, 1455–1474 (1999).
19. Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845 (2001).
20. Pfisterer, A. B. & Schmid, B. Diversity-dependent production can decrease the stability of ecosystem functioning. *Nature* **416**, 84–86 (2002).
21. Nyström, M., Folke, C. & Moberg, F. Coral reef disturbance and resilience in a human-dominated environment. *Trends Ecol. Evol.* **15**, 413–417 (2000).
22. Pandolfi, J. M. Response of Pleistocene coral reefs to environmental change over long temporal scales. *Am. Zool.* **39**, 113–130 (1999).
23. Pandolfi, J. M. *et al.* Global trajectories of the long-term decline of coral reef ecosystems. *Science* **301**, 955–958 (2003).
24. Bellwood, D. R., Hughes, T. P., Folke, C. & Nyström, M. Confronting the coral reef crisis. *Nature* **429**, 827–833 (2004).
25. Kiessling, W. & Flügel, E. in *Phanerozoic Reef Patterns* (eds Kiessling, W., Flügel, E. & Golonka, J.) 77–92 (SEPM Special Publication 72, Tulsa, 2002).
26. Bellwood, D. R. & Hughes, T. P. Regional-scale assembly rules and biodiversity of coral reefs. *Science* **292**, 1532–1535 (2001).
27. McKinney, M. L. & Oyen, C. W. Causation and nonrandomness in biological and geological time series: temperature as a proximal control of extinction and diversity. *Palaio* **4**, 3–15 (1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements I thank the DFG and the NSF for support of this research and D. Unwin, D. Lazarus, J. Alroy and M. Aberhan for comments. This is Paleobiology Database publication no. 29.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to W.K. (wolfgang.kiessling@museum.hu-berlin.de).

Stage-structured cycles promote genetic diversity in a predator–prey system of *Daphnia* and algae

William A. Nelson¹, Edward McCauley¹ & Frederick J. Wrona²

¹Ecology Division, University of Calgary, Calgary, Alberta T2N 1N4, Canada

²Water & Climate Impacts Research Centre, National Water Research Institute, University of Victoria, Victoria, British Columbia V8W 3P5, Canada

Competition theory predicts that population fluctuations can promote genetic diversity when combined with density-dependent selection^{1,2}. However, this stabilizing mechanism has rarely been tested, and was recently rejected as an explanation for maintaining diversity in natural populations of the freshwater herbivore *Daphnia pulex*³. The primary limitation of competition theory is its failure to account for the alternative types of population cycles that are caused by size- or stage-dependent population vital rates—even though such structure both explains the fluctuating dynamics of many species⁴ and may alter the outcome of competition⁵. Here we provide the first experimental test of whether alternative types of cycles affect natural selection in predator–prey systems. Using competing *Daphnia* genotypes, we show that internally generated, stage-structured cycles substantially reduce the magnitude of selection (thereby contributing to the maintenance of genetic diversity), whereas externally forced cycles show rapid competitive exclusion. The change in selection is ecologically significant, spanning the observed range in natural populations³. We argue that structured cycles reduce selection through a combination of stalled juvenile development and stage-specific mortality. This potentially general fitness-equalizing mechanism may reduce the need for strong stabilizing mechanisms to explain the maintenance of genetic diversity in natural systems.

The maintenance of genetic diversity is a central problem in biology. It is particularly acute when considering populations composed of asexual clones with extensive niche overlap. Natural populations of zooplankton, such as *Daphnia*, are a good example because they are rich in genotypic diversity^{3,6–10}, and observations of correlated dynamics in fecundity among clones^{11,12} suggest that genotypes have high resource overlap. Diversity is maintained in spite of the fact that coexisting genotypes often show strong fitness differences in the laboratory^{11,13}. These observations have been declared the second ‘paradox of the plankton’¹⁴ because they contradict the principle of competitive exclusion¹⁵, which postulates that only a single competitor should remain at equilibrium.

Competition theory offers a compelling solution to this paradox. It proposes that combining density-dependent selection with fluctuating population dynamics could allow coexistence among competitors^{1,2}. This mechanism requires that the relative fitness among competitors changes as the resource abundance changes, such that fluctuations in resource generate fluctuations in selection. Genetic diversity is thus maintained through fluctuating selection, which is considered a stabilizing mechanism that allows for coexistence by balancing fitness differences over time¹⁶. Recently, one of the first tests of this hypothesis in natural systems has been performed³ by analysing numerous time series of *Daphnia* genotype frequencies and population dynamics. No evidence was found for the maintenance of genetic diversity in *Daphnia* populations by fluctuating selection, suggesting that coexistence is more likely explained by an equalizing mechanism that reduces fitness differences among genotypes¹⁶.

Competition theory was largely developed under the assumption that populations can be modelled without considering the natural

size- and stage-dependent structure in the vital rates⁵ (for example, size-specific birth rate). However, recent work has demonstrated that structured models introduce alternative types of population cycles^{5,17,18}, which are better able to explain the range of fluctuating dynamics observed in nature^{4,19,20}. This is particularly relevant for *Daphnia*; although natural populations are dominated by stage-structured cycles¹⁹, populations can display a range of qualitatively different dynamics²¹. Here we test whether alternative types of cycles influence the rate of selection among three *Daphnia pulex* genotypes, by contrasting competition in two experimental systems. The first is a semi-chemostat environment that allows complete control of the algal resource production. We designed this experimental system to reliably generate stable, small amplitude, and large amplitude population dynamics. The second system allows resource dynamics to be set internally by population interactions in an environment that promotes stage-structured cycles. This contrast between populations under a range of externally driven algal dynamics, versus populations where the algal dynamics are internally generated by the coupled interaction between *Daphnia* and their algal prey, provides a strong test for the role of qualitatively different cycles.

We found that selection was rapid in all externally driven environments (Fig. 1). The same genotype becomes dominant in all replicates, demonstrating that competition among these competitors is deterministic. The cycling resource supply was chosen so that the predator dynamics would mimic the consumer–resource cycles observed in this predator–prey system²¹. These properties include both the period (24 days), and the amplitude (~10-fold ratio from peak to nadir) (see Fig. 1a in ref. 21 for a comparison). Interestingly, we found that these fluctuating population dynamics

actually increased the rate of selection among the three genotypes compared with the stable (constant) environments (Fig. 1). In contrast, internally generated algal dynamics strongly reduced selection among genotypes relative to any of the externally driven environments (Fig. 2). Figure 3 shows the coefficient of selection for all experiments (see Supplementary Methods for details on calculating selection). The change in selection between the different cycles is significant—both statistically ($P \ll 0.05$) and ecologically. Ecological significance can be functionally defined by considering the length of time it takes a genotype that comprises half of the population (population, $p = 0.5$) to be nearly lost from the system ($p = 0.001$); referred to as quasi-loss. Quasi-loss increased from 73 days under large amplitude, externally driven cycles to over 350 days when algal dynamics were internally generated. This change is larger than the difference between natural populations that display strong competitive exclusion within a temperate-zone growing season, and those with stable genotypic frequencies³.

Why do different dynamics produce such a pronounced change in selection? The contrast in population stage-structure between externally forced and internally generated environments provides an explanation. Internally generated cycles begin with a burst of *Daphnia* fecundity, followed by an immediate increase in the density of small juveniles (Fig. 4a). After a lag of ~13 days, there is an increase in the density of large juveniles, which may be followed by a small increase in the density of adults. This cycle is repeated again after day 40, although with a decreased amplitude (Fig. 4a). These characteristics define stage-structured cycles^{19,21}. Figure 4b shows the population structure in small amplitude, externally driven environments, which is representative of what we found for all forced environments. In contrast to stage-structured cycles,

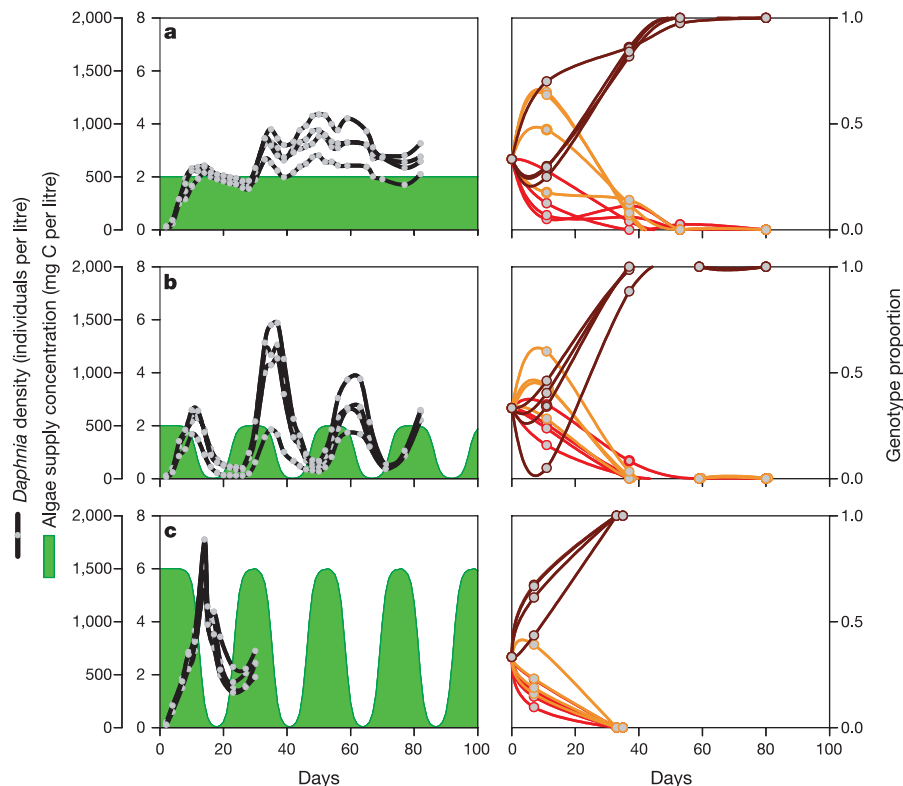


Figure 1 Population and genotype dynamics in three externally driven environments. The left panels show the population dynamics, and the right panels the corresponding genotype frequencies. Black lines are censused population dynamics and green shaded areas show the resource supply dynamics. The dark red lines are the frequency of genotype '9H', orange lines are genotype '6H', and the red lines are genotype 'RES'. Grey

circles show measured data points. Four replicates are shown for each externally driven environment. **a**, Constant resource supply. **b**, Small amplitude resource supply. **c**, Large amplitude resource supply. Experiments were terminated once the populations were entirely composed of genotype '9H'.

these are characterized by a burst of fecundity, followed by an immediate increase in the density of small juveniles, large juveniles and adults. These characteristics are the same as those found in consumer–resource cycles of *Daphnia* and algae²¹. The contrast in population structure suggests that during the growth phase of the cycle, development through the juvenile stages into adults is fast in externally driven cycles and slow in stage-structure cycles.

We propose that this change in development rate can explain the observed difference in selection. Because juvenile development is stalled during the growth phase of stage-structured cycles, most juveniles die from starvation during the decline phase without recruiting into the adult stage. Therefore, any differences among competitors in fecundity or juvenile growth are erased by mortality and failure to recruit. In contrast, the rapid development of juveniles into adults during the growth phase in externally driven environments means that any differences in fecundity or juvenile growth are carried into the adult stage. Thus, even though most juveniles die from starvation during the forced declines in externally driven environments, the differences in genotype performance are still reflected in the surviving adults. It is important to recognize that the mechanism we propose does not require any asymmetry in juvenile or adult competitive ability among genotypes. For example, both juvenile and adult mortality of the less successful competitor can be higher than the mortality rates of the more successful competitor. The change in through-stage development rate, combined with stage-specific mortality, reduces the fitness differences among genotypes, providing an equalizing mechanism for co-existence¹⁶.

The population dynamics and stage-structure in externally driven cycling environments quantitatively resembles those of large amplitude consumer–resource cycles found in coupled *Daphnia*–algal experiments. This means that the temporal pattern

of population growth rates is similar, implying that the pattern of per capita ingestion is also similar. Further evidence for this conjecture comes from the match between the amplitude and period of *Daphnia* fecundity in the externally driven experiments (Fig. 4b) and the fecundity patterns found in consumer–resource cycles (Fig. 2b in ref. 21). As genotypes only compete through the common algal prey, it is likely that selection in the driven cycling environments is representative of what would be found in coupled consumer–resource cycles.

The equalizing mechanism we have shown experimentally does not provide a means for indefinite coexistence among competitors, but it may resolve the paradox of genetic diversity in zooplankton. The expectation of rapid within-season selection among coexisting genotypes comes from two lines of evidence. The first is from laboratory life-table experiments that show significant genotype differences in clutch size, survivorship and age at maturation^{6,11,13,22,23}, and the second directly from laboratory competition experiments that show rapid selection among genotypes under conditions of frequent resource replenishment^{11,13}. Such experiments demonstrate that coexisting genotypes have the potential to show strong fitness differences. Given that natural *Daphnia* populations are dominated by stage-structured cycles¹⁹, our results suggest that population dynamics provide an equalizing mechanism that reduces selection to a point where there is little change in the genetic structure during a season. Reduced selection among strong competitors allows for weaker stabilizing mechanisms to maintain genotypic diversity over the long-term¹⁶. For example, asymmetries between asexual competition during the season and the production and survivorship of over-wintering diapause eggs becomes more important for maintaining genotypic diversity in *Daphnia*^{7,24}.

Our results demonstrate clearly that qualitatively different

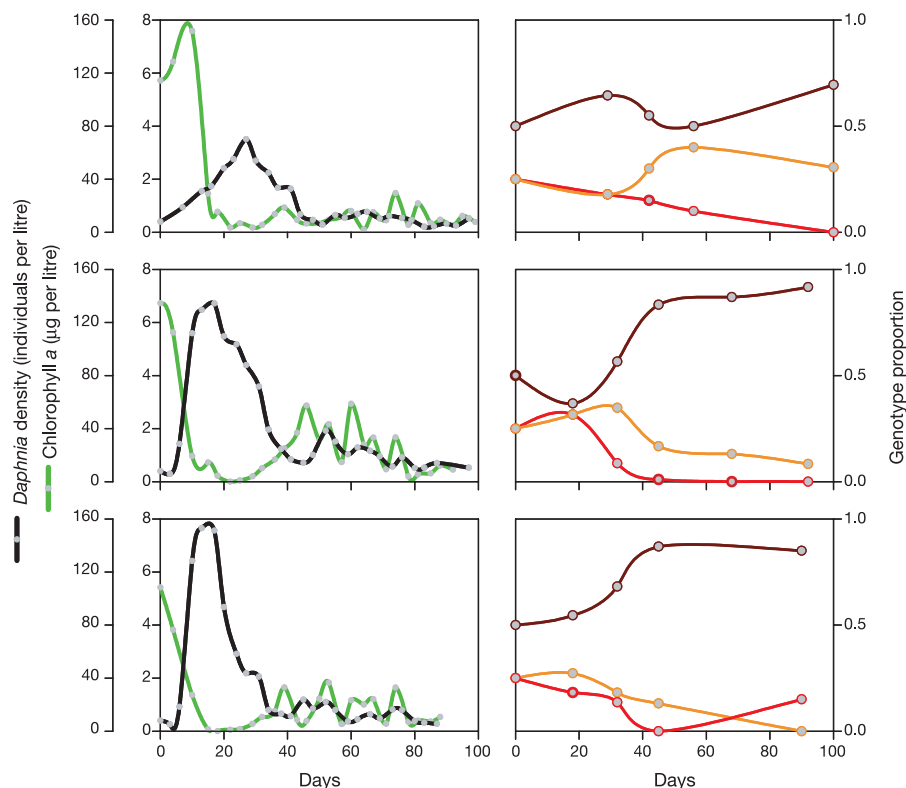


Figure 2 Population and genotype dynamics in coupled predator–prey environments. The left panels show the population dynamics, and the right panels the corresponding genotype frequencies. Black lines show the population dynamics and green lines show the

chlorophyll *a* abundance. The dark red lines show the frequency of genotype '9H', orange lines are for genotype '6H', and the red lines are for genotype 'RES'. Grey circles show the measured data points. Three representative replicate experiments are shown.

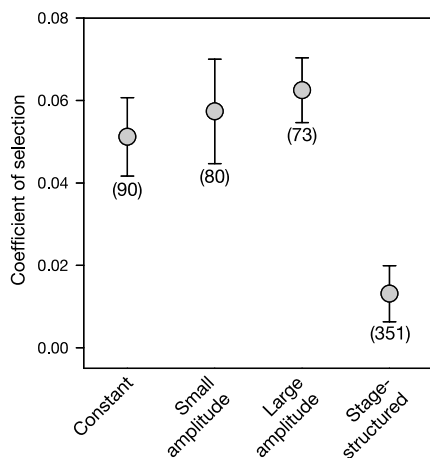


Figure 3 Selection coefficients for each treatment. Error bars show 95% confidence intervals from four replicates in the externally driven environments, and five replicates in the coupled environments. Numbers in parentheses represent the ecological significance of selection in each treatment, which is the number of days it would take for a genotype comprising half the population ($p = 0.50$) to be nearly lost ($p = 0.001$). Illustrated values are for the most successful competitor, genotype '9H'.

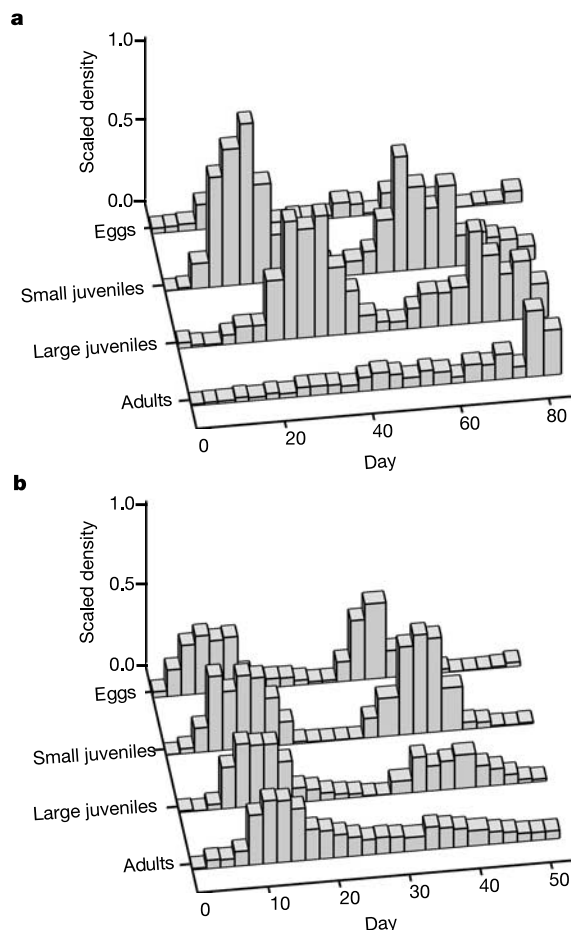


Figure 4 Population structure in coupled predator-prey environments and externally driven environments. Scaled density is shown to highlight the temporal dynamics of egg, small juvenile, large juvenile, and adult stages in each environment. Scaled density is calculated as the relative density between an exponential model fitted to the peaks of each cycle and the raw data. **a**, Internally generated, stage-structured dynamics. **b**, Small amplitude, externally driven cycles.

population dynamics have a pronounced effect on competition. Both stage-structured cycles and consumer-resource cycles emerge from a single, structured modelling framework^{17,21}. This theory argues that qualitatively different population dynamics can change the realized life history of an individual through the interplay between resource abundance, resource-dependent development rate and stage-specific mortality⁵. Our conclusions suggest that this interplay also has a strong influence on the realized selection among competitors in dynamical systems, providing a novel equalizing mechanism that may have wide applicability in natural systems. More generally, these results support the assertion that incorporating the natural size- or stage-dependent structure of individual life history into ecological and evolutionary theory leads to greater insight into natural systems⁵. □

Methods

Driven experiments

These experiments were designed to reliably generate a range of *Daphnia* population dynamics by manipulating the algal resource supply. The apparatus consists of a semi-chemostat with a double chamber design: an outer cylinder acts as a tank and an inner mesh cylinder is divided to hold four replicate zooplankton populations. The mesh size of the inner chamber is 100 μm ; this allows good algae transport but retains all *Daphnia*. An axenic food supply (cultured separately, see below) is continuously dripped into the top of each population, with the excess volume drawn out from the bottom. The flow rate is held constant at 1 litre per day in the 3-litre chamber. To manipulate prey dynamics, we set the algal concentration of the food supply each day as shown in Fig. 1. The experimental conditions were maintained at 25 °C and carried out in the dark to prevent algal growth.

We used the single-cell green algae *Chlamydomonas reinhardtii* as the prey resource. Algae were cultured in artificial pond water²⁵ and harvested at a density that ensured high-quality food. Algae abundance was estimated by preserving samples (in duplicate) in Lugol's solution²⁶ and counting 16 fields of view in a Sedgwick Rafter counting chamber. Cell counts were converted to biomass using a measured cell carbon content of 1.3345×10^{-8} mg C per cell.

Zooplankton populations were started from four adults of each genotype, using cultures that had been maintained under experimental conditions for many generations. *Daphnia* populations were completely censused every two days by gently removing all individuals from a section and measuring the size class and number of asexual or sexual eggs for each individual under a dissecting microscope²¹. This procedure allows the populations to be counted without harming the individuals²¹. Small juveniles are <1.0 mm, large juveniles are 1.0–1.4 mm, and adults are >1.4 mm in length²¹.

Coupled experiments

These experiments were conducted in 80-litre aquaria at 23 °C under a 14 h/10 h light/dark cycle. Water was collected from a local pond, filtered through a 5- μm filter and enriched with 10 $\mu\text{g l}^{-1}$ phosphorus (as KH_2PO_4) and 400 $\mu\text{g l}^{-1}$ nitrogen (as $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$ and NaNO_3). The algal inoculum was a mix of *Scenedesmus obliquus*, *Selenastrum capricornutum*, *Chlorella vulgaris* and *Chlamydomonas* spp. and was allowed to grow for five days before the inoculation of zooplankton. Zooplankton populations were started using 8.5 juveniles per litre and 2.1 adults per litre of each genotype.

The sides and bottom of all tanks were scraped daily to prevent the build-up of bacteria, and sediments were ground and sonified weekly to retain nutrients in the water column²¹. *Daphnia* abundance was estimated twice weekly by live counts of 2-litre sub-samples, using the same methods described for the driven experiments. Chlorophyll *a* was sampled at the same time by siphoning a 200 ml volume from the tank, passing it through a 35- μm mesh to remove *Daphnia*, and filtering it onto a Whatman glass micro-fibre filter (GF/C). These algae samples were frozen for fluorometric analysis at a later date²⁶.

Electrophoresis

We used three genotypes of *Daphnia pulex*, each isolated from a different body of water in Alberta, Canada. The electrophoretic identity of each genotype was determined by standard cellulose acetate protein electrophoresis²⁷. We found that the polymorphic loci phosphoglucose isomerase and lactate dehydrogenase, were sufficient to clearly distinguish among genotypes.

Genotypic frequencies were estimated less often than population dynamics to minimize sampling mortality. Frequencies were estimated biweekly in the coupled experiments by conducting electrophoresis on 1% of the population from each tank. The number of individuals assayed at each sampling time typically varied from 20 to 30, with an average of 24 depending on the population dynamics. This procedure ensured that sampling did not introduce density-dependence to the mortality rates. Genotype frequencies in the driven experiments were estimated near the peaks of each cycle by removing and conducting electrophoresis on 5% of the population, which corresponded to between 30 and 60 individuals.

Estimating selection

Here we outline the methodology used to calculate selection, but refer the reader to the Supplementary Methods and ref. 3 for full details regarding the derivation of equations

and development of statistical methods. Briefly, the statistical model is:

$$\frac{N_i(t)}{N_T(t)} = f_i(t) \quad (1)$$

where $N_i(t)$ are the observed counts for genotype i in a sample of size $N_T(t)$ at time t , and $f_i(t)$ is a quadratic function representing the temporal genotype dynamics. We use a Dirichlet-multinomial distribution as the error structure for the observed counts to accommodate possible over-dispersion in the data. Equation (1) was fitted with the statistical library VGAM²⁸ in the R software environment (R Development Core Team, <http://www.R-project.org>). The temporal dynamics of selection are calculated from

$$s(t)_i = \frac{dp(t)_i}{dt} \frac{1}{p(t)_i(1 - p(t)_i)} \quad (2)$$

where $s(t)_i$ is the coefficient of selection for genotype i at time t , and $p(t)_i$ is the estimated relative abundance. Net selection is calculated from equation (2) as:

$$\bar{s}_{ni} = \frac{\int_{t_1}^{t_2} s_i(t) dt}{(t_2 - t_1)} \quad (3)$$

Received 20 September; accepted 12 November 2004; doi:10.1038/nature03212.

1. Armstrong, R. A. & McGehee, R. Competitive exclusion. *Am. Nat.* **115**, 151–170 (1980).
2. Huisman, J. & Weissing, F. J. Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999).
3. Nelson, W. A. *Competition in Structured Zooplankton Populations: Coupling Population Genetics and Dynamics Using Theoretical and Experimental Approaches*. Ph.D. Thesis, Univ. Calgary (2004).
4. Murdoch, W. W. *et al.* Single-species models for many-species food webs. *Nature* **417**, 541–543 (2002).
5. de Roos, A. M., Persson, L. & McCauley, E. The influence of size-dependent life-history traits on the structure and dynamics of populations and communities. *Ecol. Lett.* **6**, 473–487 (2003).
6. Lynch, M. The limits to life history evolution in *Daphnia*. *Evolution* **38**, 465–482 (1984).
7. Lynch, M. The consequences of fluctuating selection for isozyme polymorphisms in *Daphnia*. *Genetics* **115**, 657–669 (1987).
8. Mort, M. A. Bridging the gap between ecology and genetics: the case of freshwater zooplankton. *Trends Ecol. Evol.* **6**, 41–45 (1991).
9. Dufresne, F. & Hebert, P. D. N. Polyploidy and clonal diversity in an arctic cladoceran. *Heredity* **75**, 45–53 (1995).
10. Hebert, P. D. N. & Crease, T. J. Clonal diversity in populations of *Daphnia pulex* reproducing by obligate parthenogenesis. *Heredity* **51**, 353–369 (1983).
11. Weider, L. J. Spatial and temporal genetic heterogeneity in a natural *Daphnia* population. *J. Plankton Res.* **7**, 101–123 (1985).
12. Hebert, P. D. N. Niche overlap among species in the *Daphnia carinata* complex. *J. Anim. Ecol.* **46**, 399–409 (1977).
13. Loaring, J. M. & Hebert, P. D. N. Ecological differences among clones of *Daphnia pulex* Leydig. *Oecologia* **51**, 162–168 (1981).
14. Hebert, P. D. N. & Crease, T. J. Clonal coexistence in *Daphnia pulex* (Leydig): another planktonic paradox. *Science* **207**, 1363–1365 (1980).
15. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292–1298 (1960).
16. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 334–366 (2000).
17. de Roos, A. M. & Persson, L. Competition in size-structured populations, mechanisms inducing cohort formation and population cycles. *Theor. Popul. Biol.* **63**, 1–16 (2003).
18. McCauley, E., Nisbet, R. M., de Roos, A. M., Murdoch, W. W. & Gurney, W. S. C. Structured population models of herbivorous zooplankton. *Ecol. Monogr.* **66**, 479–501 (1996).
19. McCauley, E. & Murdoch, W. W. Cyclic and stable populations: plankton as paradigm. *Am. Nat.* **129**, 97–121 (1987).
20. Persson, L., Byström, P., Wahlström, E., Andersson, J. & Hjelm, J. Interactions among size-structured populations in a whole lake experiment: size- and scale dependent processes. *Oikos* **87**, 139–156 (1999).
21. McCauley, E., Nisbet, R. M., Murdoch, W. W., de Roos, A. M. & Gurney, W. S. C. Large-amplitude cycles of *Daphnia* and its algal prey in enriched environments. *Nature* **402**, 653–656 (1999).
22. Spitze, K. *Chaoborus* predation and the life-history evolution in *Daphnia pulex*: temporal pattern of population diversity, fitness, and mean life-history. *Evolution* **45**, 82–92 (1991).
23. Carvalho, G. R. The clonal ecology of *Daphnia magna* (Crustacea: Cladocera) II: thermal differentiation among seasonal clones. *J. Anim. Ecol.* **56**, 469–478 (1987).
24. Mitchell, S., Carvalho, G. R. & Weider, L. J. Stability of genotype frequencies in an intermittent *Daphnia magna* population. *Arch. Hydrobiol. Spec. Issues Adv. Limnol.* **52**, 185–194 (1998).
25. Kilham, S. S., Kreeger, D. A., Lynn, S. G., Goulder, C. E. & Herrera, L. COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia* **377**, 147–159 (1998).
26. Watson, S., McCauley, E. & Downing, J. A. Variation in algal community structure with enrichment. *Can. J. Fish. Aquat. Sci.* **49**, 2605–2610 (1992).
27. Hebert, P. D. N. & Beaton, M. J. *Methodologies for Allozyme Analysis using Cellulose Acetate Electrophoresis* (Helena Laboratories, Beaumont, Texas, 1993).
28. Yee, T. W. & Wild, C. J. Vector generalized additive models. *J. Roy. Stat. Soc. B* **58**, 481–493 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Hebert for his electrophoresis expertise and J. Fox for constructive comments on the original manuscript. The research was supported by grants from NSERC to E.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.A.N. (wanelson@ucalgary.ca).

Host immunity and synchronized epidemics of syphilis across the United States

Nicholas C. Grassly, Christophe Fraser & Geoffrey P. Garnett

Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK

A central question in population ecology is the role of ‘exogenous’ environmental factors versus density-dependent ‘endogenous’ biological factors in driving changes in population numbers¹. This question is also central to infectious disease epidemiology, where changes in disease incidence due to behavioural or environmental change must be distinguished from the nonlinear dynamics of the parasite population². Repeated epidemics of primary and secondary syphilis infection in the United States over the past 50 yr have previously been attributed to social and behavioural changes³. Here, we show that these epidemics represent a rare example of unforced, endogenous oscillations in disease incidence, with an 8–11-yr period that is predicted by the natural dynamics of syphilis infection, to which there is partially protective immunity⁴. This conclusion is supported by the absence of oscillations in gonorrhoea cases, where a protective immune response is absent^{5,6}. We further demonstrate increased synchrony of syphilis oscillations across cities over time, providing empirical evidence for an increasingly connected sexual network in the United States.

Syphilis and gonorrhoea are endemic sexually transmitted diseases in the United States. After the introduction of widespread screening and treatment by the Surgeon General Thomas Parran after the Second World War, reported cases of primary and secondary syphilis dropped to an all-time low in 1955 before they rose again and began to oscillate with an approximately 10-year cycle (Fig. 1a). The introduction of treatment for gonorrhoea also resulted in a decline in incidence during the 1950s, but, unlike syphilis, periodic oscillations did not ensue (Fig. 1b). Oscillations in the incidence of syphilis have been attributed to social and behavioural changes³. For example, the epidemic in the 1970s is attributed to the sexual revolution and gay liberation movement, the epidemic in the 1980s to poverty, urbanization and crack-cocaine-related prostitution. Here we re-examine this idea by gathering and analysing data on syphilis and gonorrhoea cases reported to the US Centers for Disease Control and Prevention between 1941 and 2002 for 68 cities across the United States. These data provide a rare opportunity to examine the relative roles of endogenous and exogenous factors in driving infectious disease dynamics, with the availability of both syphilis and gonorrhoea case reports allowing a comparative approach to testing hypotheses.

We tested the time series of primary and secondary syphilis and gonorrhoea case reports from individual cities for significant periodicity by calculating smoothed periodograms using standard spectral analysis techniques⁷. We focus on the period 1960–93, beginning with the initial resurgence of infection after the introduction of treatment and ending before declines due to the HIV epidemic, and thus avoid complications that can arise from non-stationarity in the epidemic period (in which case wavelet approaches may be more appropriate⁸). The analysis confirms the oscillatory nature of endemic syphilis previously noted from national case numbers (Fig. 1c). The resolution of the period estimate is restricted by the number of years of observations, but the average spectrum for these cities peaks at about 8–10 yr and a significant periodicity is found between 8.25 and 11 yr for 21 of the 47 cities with complete reporting during 1960–93 ($P < 0.05$).

Analysis of gonorrhoea case reports from the same cities reveals flat spectra, indicating an absence of any periodicity in the data over the scale of interest (Fig. 1d) (weak seasonality has been demonstrated for gonorrhoea, but not syphilis, dynamics possibly due to a small increase in rates of sexual partner change during the summer⁹). The contrasting patterns seen for syphilis and gonorrhoea time series and their spectra are illustrated for the largest and smallest cities with complete data from the northeast and south of the United States in Fig. 2a.

Primary and secondary syphilis and gonorrhoea both have a relatively short duration of infection (<6 months)^{10,11} and similar transmission probabilities and case detection and treatment protocols^{12,13}. This predicts a close relationship between levels of unsafe sex and disease incidence¹⁴, and indeed these diseases typically infect similar 'core' groups¹⁵. If social and behavioural changes are responsible for the oscillations observed for primary and secondary syphilis infections, then similar oscillations should be observed for gonorrhoea. Their absence leads us to an alternative explanation based on a key biological difference between syphilis and gonorrhoea infections.

In both treated and untreated syphilis partial immunity to reinfection occurs, either due to acquired immune memory or continued presence of antigens respectively⁴. The presence of immunity has important effects on the dynamics of infectious diseases. When the period of infectiousness is short compared with immunity, fluctuations in incidence occur over a timescale determined by the supply of susceptible individuals¹⁶. This can be illustrated by a simple SIRS compartmental model with 'susceptible', 'infected' and 'recovered' (immune) states followed by the loss of immunity and a return to the susceptible state. The stochastic version of this model shows sustained oscillations in incidence, with a period that is a function of the natural history of the infection and the rate at which susceptible individuals enter the population, but independent of population size (see Methods). These oscillations

persist even if a significant fraction of infected individuals do not develop protective immunity (see Supplementary Fig. 1).

For parameters consistent with the natural history of syphilis, the SIRS model predicts both the period and amplitude of oscillations in incidence seen for US cities (Fig. 2b). The basic reproductive number (R_0) of an infection can be defined as the average number of secondary infections caused by a single infectious individual in a completely susceptible population, and is a key parameter determining disease invasion, spread and persistence. A range of values for R_0 , typically less than about 3, are consistent with the observed 8–11-yr oscillations when realistic rates of treatment and entry to the susceptible population are specified (Fig. 2c). A small R_0 for syphilis in the United States is also indicated by the slow increase in the number of cases observed during syphilis outbreaks and estimates based on transmission probabilities and rates of sexual partner change¹⁰.

The close consistency of both period and amplitude of oscillations in syphilis incidence in the model and data suggests that it is the endogenous nonlinear dynamics of syphilis after the introduction of treatment that are largely responsible for the oscillations in incidence rather than exogenous forcing by repeated social and behavioural change. This conclusion is supported by comparison with gonorrhoea, which does not lead to any protective immunity, as revealed by repeat infections with the same serovars seen for both STD clinic patients⁵ and inoculated human male volunteers⁶. Gonorrhoea dynamics can therefore be described by an SIS model of incidence, where the recovered state is absent, and which does not predict oscillations in incidence. Oscillations are only possible when a period of immunity (or death) removes individuals from the population of susceptibles. The gradual rise and fall in gonorrhoea incidence since 1960 is therefore perhaps more likely to reflect the underlying dynamics of unsafe sexual behaviour in the United States (although case numbers may also be affected by the effort put into case detection and reporting).

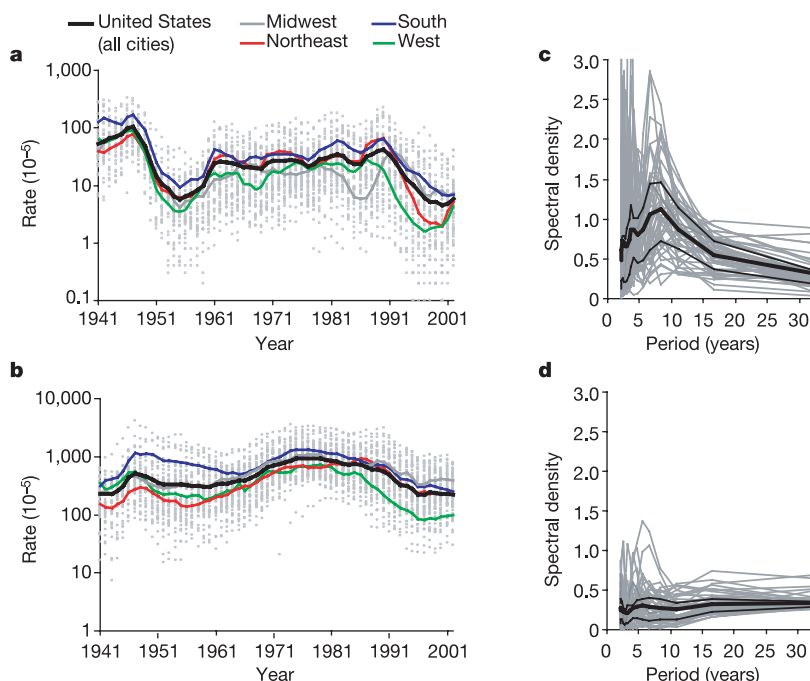


Figure 1 Syphilis and gonorrhoea cases in the United States. **a, b**, Case reports for primary and secondary syphilis (**a**) and gonorrhoea (**b**) plotted against time as grey points for each of the 68 major cities of the United States. Cases are plotted as a rate per 100,000 of the city population on a log₁₀ scale and regional and national totals are shown

as lines. **c, d**, The spectra of the differenced data for syphilis (**c**) and gonorrhoea (**d**) are shown for 47 cities with complete reporting over the period 1960–93. The average spectrum is shown by the thick black line and inter-quartile range by the thin black lines.

The reported cases of syphilis come from a network of cities, which are connected by travel and sexual contact between individuals from different cities. This broader sexual network may also be important for the dynamics of syphilis. Behaviour and treatment patterns are likely to vary between cities, resulting in different R_0 values and different intrinsic periods for oscillations in incidence. However, coupling through contact between individuals from different populations can synchronize both the period and phase of oscillations in disease incidence¹⁷.

We demonstrate the importance of coupling for primary and secondary syphilis dynamics by calculating the correlation in incident cases reported for the different cities. Shortly after the introduction of treatment the average correlation between cities is low, but during the latter half of the twentieth century it shows a large increase (Fig. 3a). Peaks in correlation typically occur during epidemics, as the disease sweeps through the susceptible population in different cities. Over time these peaks have increased in size, with an average correlation coefficient over the period 1980–96 of 0.24 compared with 0.04 during 1960–79. This suggests an increasingly connected sexual network across the United States and is certainly consistent with the well documented increases in passenger travel over the last few decades¹⁸.

The pattern of increase in correlation of incidence during epidemics is well captured by a simple metapopulation version of the stochastic SIRS model, where a fraction ε of sexual contacts are 'global' rather than local (Fig. 3b). In this model oscillations in incidence synchronized in phase and frequency emerge as the level

of coupling increases. Synchronization is balanced by stochastic drift in the phase of the oscillations in different cities and local extinction followed by re-introduction of infection. For a reasonable number of populations, increased coupling is found to increase the average correlation of incidence in the metapopulation at all levels of coupling (Fig. 3c). This contrasts with the threshold effects observed for deterministic models¹⁷.

Cross-correlation is significantly lower among the ten smallest cities compared with the ten largest cities ($P < 0.001$), even after accounting for sampling error, as expected if random drift in phase and local extinction are important for the dynamics of the disease. However, populations at risk of infection in these smaller cities may also be less strongly coupled to other cities. This has been demonstrated for several directly transmitted infections, including HIV, which have spread predominantly through the size hierarchy of cities^{19,20}. In general the extent of coupling required to synchronize disease dynamics in different populations will depend on the exact nature of the connections between populations^{21,22}.

Syphilis incidence in the United States is concentrated among African Americans from lower socioeconomic groups in the south and among men who report sex with other men²³. The epidemiological linkage of these groups predicts endogenous oscillations in incidence, which are coupled in phase and frequency as for the metapopulation. More generally, SIR infection dynamics in structured populations described by an explicit contact network show periodic oscillations for all but the lowest level of network disorder²⁴. Social and behavioural changes among these groups will, of

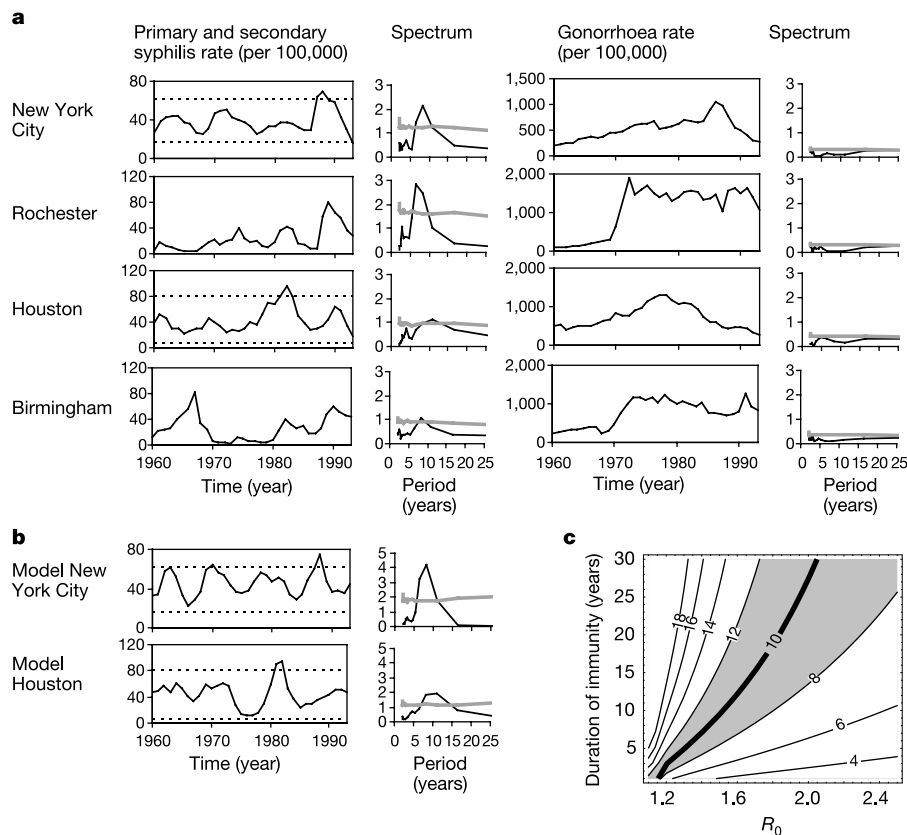


Figure 2 Oscillations in syphilis incidence due to immunity. **a**, Time series of syphilis and gonorrhoea case reports together with their spectra for the largest and smallest cities in the northeast (New York City, Rochester) and south (Houston, Birmingham). **b**, Stochastic realizations of a SIRS model with R_0 , population size N and the duration of immunity estimated for New York City and Houston, respectively. The estimated amplitude (± 2 s.d., dashed lines) and period (spectral density, black lines) of these oscillations in incidence is

the same for the model and data. Significant periodicity may be assessed by comparing the spectral density with the 95th percentile of the spectra of the randomized data (grey line). **c**, The period of the oscillations for different R_0 values and duration of immunity, in the case that syphilis infection lasts 2 months on average before treatment and individuals are sexually active for an average of 30 yr.

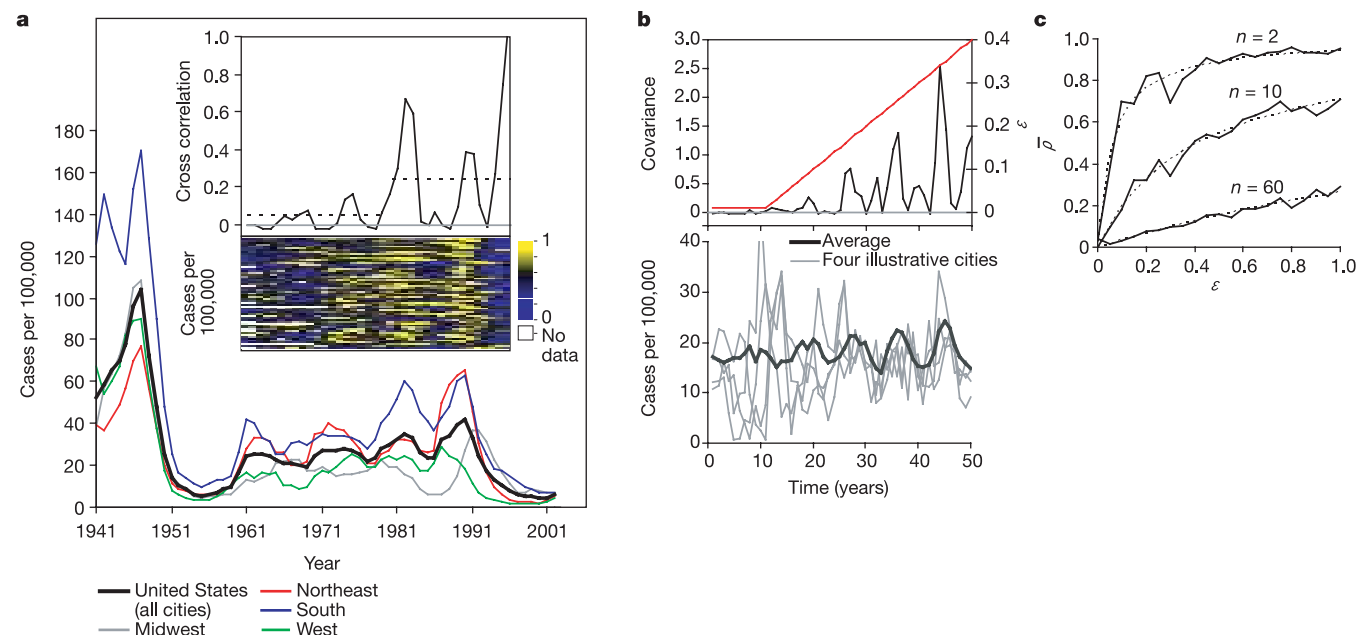


Figure 3 The metapopulation dynamics of syphilis. **a**, Primary and secondary syphilis case reports per 100,000 population by region and for 59 cities as a colour intensity plot (inset) for the period 1960–96 together with the average cross correlation. Reported rates in the intensity plot are scaled between 0 and 1 after transforming by $\log_{10}(1 + \text{rate})$, and cities are arranged in descending order of population size (from top to bottom). Vertical bands corresponding to synchronized epidemics begin to emerge to the right of the plot. Correlation of case reports between cities given by standardized covariance (solid line) is shown for each year, and the average correlation coefficient ($\bar{\rho}$) (dashed line) for the

periods 1960–79 and 1980–96. **b**, Stochastic realization of a metapopulation SIRS model with 60 populations showing the standardized covariance (black line) over time as coupling (ε ; red line) increases linearly from 0.01 to 0.4. **c**, Average correlation ($\bar{\rho}$, solid line) between n populations for numerical simulations of the stochastic metapopulation SIRS model at different levels of coupling. The relationship is well approximated by $\bar{\rho} \approx \varepsilon / (\xi + \varepsilon)$ (dashed line) where ξ is a parameter that can be estimated directly from the simulations (compare with moment closure approximation for the two-population case²¹).

course, affect the incidence of syphilis, even if they are not responsible for the large amplitude oscillations in case reports. In particular, the epidemic in the late 1980s was uniquely associated with a significant increase in the fraction of infections among women, probably the result of an increase in crack-cocaine-related prostitution³, and the decline of both syphilis and gonorrhoea throughout the 1990s a result of mortality and behaviour change due to the HIV epidemic²⁵. In general, deviations from an oscillatory pattern may be driven by changes in behaviour, surveillance systems or random local extinction of infection and phase drift. Disentangling these factors remains a formidable challenge.

Here, we have demonstrated large amplitude oscillations in syphilis incidence, absent for gonorrhoea, which can be explained by the endogenous dynamics of syphilis where there is a partially protective host immune response. This important role of host immunity driving endogenous disease dynamics highlights the need to carefully interpret data from routine surveillance and intervention studies, where changes in incidence may not always be attributable to changes in behaviour or the environment. In particular, recent synchronized increases in syphilis incidence across the United States have sparked concern that there is growing complacency about the risks of HIV and returns to unsafe sexual behaviour, particularly among men who report sex with other men^{26,27}. Our analyses, however, suggest that these synchronized increases may largely reflect natural dynamics after a period of build up of non-immune individuals, a conclusion supported by the lack of any correlated increases in gonorrhoea case reports. □

Methods

Data

We compiled reported cases and rates of primary and secondary syphilis and gonorrhoea infection for the period 1941–2002 for 68 major cities (population size >200,000) of the United States, from data collected by the Centers for Disease Control and Prevention.

Spectral analysis

Long-term trends in the mean reported rate for each city were removed before spectral analysis by subtracting a best-fit straight line and the series normalized to zero mean and unit variance. Periodograms representing Fourier transforms of the differenced data ($\nabla c_t = c_t - c_{t-1}$ where c_t is the observed case report rate at time t) were calculated and smoothed using a standard Daniell window of width 3 yr². Differencing removes any long-term changes in the mean rate (non-stationarity) remaining after the linear transformation, resulting in a suppressed spectral density for low frequency ‘noise’ (see Supplementary Methods).

Tests for significant periodicity in each city spectrum were based on the spectra of 1,000 random permutations of the differenced data for each city. Randomization removes any periodicity and is equivalent to the hypothesis of random noise in the differenced data.

Cross-correlation

Correlation of the rate of primary and secondary syphilis infection between 59 cities was estimated using the average of Pearson’s correlation coefficient ($\bar{\rho}$) between each city pair (nine cities with 25 or less years of reporting were removed from the analysis). To examine temporal trends in cross-correlation we simply plot the annual components of the average covariance, standardized by dividing by the variance across cities at each point in time to avoid spurious trends in cross-correlation that may result from changes in the variance. Unlike the average correlation, the standardized covariance does not have a natural scale between -1 and 1 , and for both measures there is no simple statistical test for significance owing to the non-independence of the correlation coefficients²⁸. We therefore make comparisons of $\bar{\rho}$ with the same measure of correlation in the metapopulation SIRS model (Fig. 3c).

The spread of disease through the size hierarchy of US cities confounds the use of spatial correlograms²⁹ (see Supplementary Methods). Instead we compare the number of significant pairwise correlations in incident cases during 1960–96 in the ten largest and ten smallest cities. Significance was assessed after bootstrapping the case report data from the ten largest cities with additional sampling error described by the binomial distribution where the number of reported cases equalled that for the smallest cities (see Supplementary Methods).

The model

We extend the classic SIRS model of infection²⁹ to allow for a network of connected populations (metapopulation). In this simple model, infection is followed by recovery (with rate ν) to an immune class with the possibility of subsequent loss of immunity (with rate γ) to return to the susceptible class. We denote the population size and number of infected individuals in population i by N_i and Y_i respectively, and prevalence by $y_i = Y_i/N_i$. The population size is equivalent to those at risk of infection, which in the case of syphilis can be considered analogous to the core group of sexually active individuals¹⁴.

We do not explicitly consider heterogeneity in sexual activity, although the results are robust to consideration of contact networks and stratification of the population (see Supplementary Methods).

The hazard of infection for susceptible individuals in the i th population is

$$\lambda_i = (1 - \varepsilon)\beta_i y_i + \varepsilon\beta_i \bar{y} \quad (1)$$

where ε is the fraction of contacts that are global, β_i the transmission parameter for population i , and $\bar{y}$ is the prevalence among global contacts, which is simply the average prevalence over the n populations of the metapopulation weighted by the transmission parameter β_i to ensure numbers of contacts between populations balance ($\bar{y} = \sum_n Y_i \beta_i / \sum_n N_i \beta_i$). This type of representation of coupling between populations has been shown to be a good approximation to more explicit mechanistic models of migration, where individuals spend periods outside their 'home' city^{21,22}.

The stochastic version of this model can give sustained oscillations in prevalence and incidence due to continued perturbation of the system by random events¹⁶. In the case that $\varepsilon = 0$, prevalence in each population oscillates with period

$$T = 2\pi / \sqrt{s(\gamma + \mu) - 0.25(\gamma + \mu + q)^2} \quad (2)$$

where μ is the birth/death rate, $q = (\gamma + \mu)/(\gamma + \mu + \nu)$, $s = (\nu + \mu)(R_0 - 1)$ and R_0 the basic reproductive number given by $R_0 = \beta/(\nu + \mu)$. We omit the subscript i for notational clarity. Although seasonality has not been shown to be important for syphilis⁹, we note that for parameters consistent with syphilis natural history the period predicted by equation (2) does not change with the inclusion of seasonal forcing, which merely acts to ensure that the period is an integer number of years. The amplitude of oscillations is given by the variance in the number of infected individuals, which can be estimated through a diffusion approximation for reasonably large N as $\sigma_y^2 = N(R_0 - 1)/R_0^2$ (ref. 29). These oscillations can also occur in the SIR model and persist in both models even when a significant fraction of individuals do not develop immunity (see Supplementary Fig. 1). As the metapopulation becomes increasingly coupled ($\varepsilon \rightarrow 1$), these oscillations become synchronized in phase and frequency, until in the fully coupled system synchronized oscillations occur with period approximated by equation (2) with a basic reproductive number given by $\bar{R}_0 + \sigma_{R_0}^2/\bar{R}_0$, where $\bar{R}_0$ is the mean and $\sigma_{R_0}^2$ the variance of the reproductive number for each population.

The model parameters γ , R_0 and N for New York City and Houston in Fig. 2b were estimated from the period, amplitude and mean of the oscillations in case reports, by making the simplifying assumption that the number of reported cases $C \approx fY$, where f is the fraction of incident infections that are reported. For this illustration we assume that the other model parameters are fixed ($f = 0.5$, $\nu = 9$ in New York City and 6 in Houston, and $\mu = 0.03$).

The impact of different levels of coupling (ε) on cross-correlation of incidence (Fig. 3b, c) was examined numerically for a metapopulation where R_0 varies across cities according to a lognormal distribution (a few cities have high R_0), with mean 2 and variance 0.5 (although the results are closely consistent for different values of the variance, including zero). The distribution of N was assumed to follow a power law, in agreement with the decennial census estimates of the US city sizes in 1970 (ref. 30), but only a small fraction (1–5%) of the total population is assumed to be at risk of infection. The average cross-correlation ρ shown in Fig. 3c is for 5,000-yr simulations. All other model parameters are fixed ($\nu = 6$, $\mu = 0.03$, $\gamma = 0.05$).

Received 30 June; accepted 30 September 2004; doi:10.1038/nature03072.

1. Bjornstad, O. N. & Grenfell, B. T. Noisy clockwork: Time series analysis of population fluctuations in animals. *Science* **293**, 638–643 (2001).
2. Pascual, M., Rodo, X., Ellner, S. P., Colwell, R. & Bouma, M. J. Cholera dynamics and el Niño–Southern Oscillation. *Science* **289**, 1766–1769 (2000).
3. Nakashima, A. K., Rolf, R. T., Flock, M. L., Kilmarx, P. & Greenspan, J. R. Epidemiology of syphilis in the United States, 1941–1993. *Sex. Transm. Dis.* **23**, 16–23 (1996).
4. Morgan, C. A., Lukehart, S. A. & Van Voorhis, W. C. Protection against syphilis correlates with specificity of antibodies to the variable regions of *Treponema pallidum* repeat protein K. *Infect. Immun.* **71**, 5605–5612 (2003).
5. Fox, K. K. et al. Longitudinal evaluation of serovar-specific immunity to *Neisseria gonorrhoeae*. *Am. J. Epidemiol.* **149**, 353–358 (1999).
6. Schmidt, K. A. et al. Experimental gonococcal urethritis and reinfection with homologous gonococci in male volunteers. *Sex. Transm. Dis.* **28**, 555–564 (2001).
7. Chatfield, C. *The Analysis of Time Series: an Introduction* 6th edn (Chapman & Hall/CRC, Boca Raton, 2003).
8. Grenfell, B. T., Bjornstad, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716–723 (2001).
9. Hethcote, H. W. & Yorke, J. A. *Gonorrhea Transmission Dynamics and Control in Lecture Notes Biomath* no. 56 (Springer, Berlin, 1984).
10. Garnett, G. P., Aral, S. O., Hoyle, D. V., Cates, W. & Anderson, R. M. The natural history of syphilis: Implications for the transmission dynamics and control of infection. *Sex. Transm. Dis.* **24**, 185–200 (1997).
11. Garnett, G. P., Mertz, K. J., Finelli, L., Levine, W. C. & St Louis, M. E. The transmission dynamics of gonorrhoea: modelling the reported behaviour of infected patients from Newark, New Jersey. *Phil. Trans. R. Soc. Lond. B* **354**, 787–797 (1999).
12. Hook, E. W. et al. Delayed presentation to clinics for sexually transmitted diseases by symptomatic patients—A potential contributor to continuing STD morbidity. *Sex. Transm. Dis.* **24**, 443–448 (1997).
13. Greenberg, J. B. et al. Learning from clients: An opportunity for sexually transmitted disease programs. *J. Public Health Manage Pract.* **8**, 59–68 (2002).
14. Wasserheit, J. N. & Aral, S. O. The dynamic topology of sexually transmitted disease epidemics: implications for prevention strategies. *J. Infect. Dis.* **174** (suppl. 2), S201–S213 (1996).
15. Hamers, F. F. et al. Syphilis and Gonorrhea in Miami—Similar clustering, different trends. *Am. J. Public Health* **85**, 1104–1108 (1995).

16. Bailey, N. T. J. *The Mathematical Theory of Infectious Diseases and its Applications* 2nd edn (Griffin, London, 1975).
17. Lloyd, A. L. & May, R. M. Spatial heterogeneity in epidemic models. *J. Theor. Biol.* **179**, 1–11 (1996).
18. U.S. Department of Transportation Bureau of Transportation Statistics. *National Transportation Statistics* 2003 (U.S. Government Printing Office, Washington DC, 2003).
19. Cliff, A. D., Haggett, P. & Smallman-Raynor, M. in *Deciphering Global Epidemics: Analytical Approaches to the Disease Records of World Cities, 1888–1912* (eds Baker, A. R. H., Dennis, R. & Holdsworth, D.) (Cambridge Univ. Press, Cambridge, 1998).
20. Wallace, R., Huang, Y. S., Gould, P. & Wallace, D. The hierarchical diffusion of AIDS and violent crime among US metropolitan regions: Inner-city decay, stochastic resonance and reversal of the mortality transition. *Soc. Sci. Med.* **44**, 935–947 (1997).
21. Keeling, M. J. & Rohani, P. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecol. Lett.* **5**, 20–29 (2002).
22. Lloyd, A. L. & Jansen, V. A. A. Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. *Math. Biosci.* **188**, 1–16 (2004).
23. Aral, S. O. The social context of syphilis persistence in the southeastern United States. *Sex. Transm. Dis.* **23**, 9–15 (1996).
24. Kuperman, M. & Abramson, G. Small world effect in an epidemiological model. *Phys. Rev. Lett.* **86**, 2909–2912 (2001).
25. Chesson, H. W., Dee, T. S. & Aral, S. O. AIDS mortality may have contributed to the decline in syphilis rates in the United States in the 1990s. *Sex. Transm. Dis.* **30**, 419–424 (2003).
26. Kahn, R. H., Heffelfinger, J. D. & Berman, S. M. Syphilis outbreaks among men who have sex with men—A public health trend of concern. *Sex. Transm. Dis.* **29**, 285–287 (2002).
27. Wolitski, R. J., Valdiserri, R. O., Denning, P. H. & Levine, W. C. Are we headed for a resurgence of the HIV epidemic among men who have sex with men? *Am. J. Public Health* **91**, 883–888 (2001).
28. Bjornstad, O. N., Ims, R. A. & Lambin, X. Spatial population dynamics: analyzing patterns and processes of population synchrony. *Trends Ecol. Evol.* **14**, 427–432 (1999).
29. Näsell, I. Stochastic models of some endemic infections. *Math. Biosci.* **179**, 1–19 (2002).
30. Gibson, C. *Population of the 100 Largest Cities and Other Urban Places in the United States: 1790–1990* (Working Paper No. 27, Population Division, US Census Bureau, Washington DC, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements The authors are grateful to M. Flock and S. Berman at the US Centers for Disease Control and Prevention for making available the disease case reports and for discussion on data quality. We also thank J. Truscott, J. Lewis, P. White, N. Ferguson, S. Riley and O. Bjornstad for advice and help in analysis of periodic time-series data. N.C.G. and C.F. would like to thank the Royal Society, and G.P.G. and C.F. the Medical Research Council for funding.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.C.G. (n.grassly@imperial.ac.uk).

How the Venus flytrap snaps

Yoël Forterre¹, Jan M. Skotheim^{2,4}, Jacques Dumais³ & L. Mahadevan^{3,4}

¹IUSTI CNRS, Université de Provence, 5 rue Enrico Fermi, 13453 Marseille Cedex 13, France

²Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WA, UK

³Department of Organismic and Evolutionary Biology, Harvard University Biological Laboratories, 16 Divinity Avenue, and

⁴Division of Engineering and Applied Sciences, Harvard University, Pierce Hall, 29 Oxford Street, Cambridge, Massachusetts 02138, USA

The rapid closure of the Venus flytrap (*Dionaea muscipula*) leaf in about 100 ms is one of the fastest movements in the plant kingdom. This led Darwin to describe the plant as “one of the most wonderful in the world”¹. The trap closure is initiated by the mechanical stimulation of trigger hairs. Previous studies^{2–7} have focused on the biochemical response of the trigger hairs to stimuli and quantified the propagation of action potentials in the leaves. Here we complement these studies by considering the post-stimulation mechanical aspects of Venus flytrap closure. Using high-speed video imaging, non-invasive microscopy techniques and a simple theoretical model, we show that the fast closure of the trap results from a snap-buckling instability, the onset of which is controlled actively by the plant. Our study identifies an ingenious solution to scaling up movements in non-

muscular engines and provides a general framework for understanding nastic motion in plants.

Plants are not known for their ability to move quickly. Nevertheless, rapid plant movements are involved in essential functions such as seed and pollen dispersal (exploding fruits in *Impatiens*, squirting cucumber and trigger plants), defence (sensitive mimosa) and nutrition (Venus flytrap, *Aldrovanda vesiculosa*, bladderwort). Of these spectacular examples that have long fascinated scientists, the leaves of the Venus flytrap (Fig. 1a), which snap together in a fraction of second to capture insects, have long been a paradigm for study; however, the mechanism by which this engine works remains poorly understood^{1,5,8–13}. The most frequently proposed explanations are an irreversible, acid-induced wall loosening⁸, and a rapid loss of turgor pressure in 'motor cells'¹³. However, the validity of both mechanisms has recently been questioned^{5,11} on the grounds that these cellular mechanisms alone cannot explain the rapidity of closure of the entire leaf on a macroscopic scale; this has led to the suggestion⁵ that elastic deformations might be important.

Any mechanistic explanation requires an understanding of the geometry of snapping. Therefore, we first quantified the change in leaf geometry during closure by painting sub-millimetric ultraviolet-fluorescent dots on the external face of the leaves and filmed closure under ultraviolet light, using high speed video at 400 frames per second (Fig. 1b, see Supplementary Methods for a movie). Using a pair of mirrors to record stereo images, we reconstructed the leaf geometry and the change therein using triangulation (Fig. 1b, c; see Methods). As Darwin had already noted¹, the leaf is curved outward

(convex) in the open state and curved inward (concave) in the closed state (Fig. 1a). The leaf shape can be naturally characterized in terms of its spatially averaged mean curvature (κ_m) and its spatially averaged gaussian curvature (κ_g), both of which are invariant under rigid body motions and are thus indicators of shape. In Fig. 1d we plot κ_m as a function of time and observe that the snapping motion is characterized by three phases: a slow initial phase (20% of total displacement in 1/3 s), a rapid intermediate phase (60% of total displacement in 1/10 s) and finally a second slow phase (20% of total displacement in 1/3 s). The existence of the three phases is consistently observed, but the quantitative values may vary. Most of the leaf displacement occurs in the intermediate phase, during which the leaf geometry changes from convex to concave. Figure 1e shows κ_g as a function of time. We see that κ_g is not constant, and also that κ_g changes slowly and then rapidly as it passes through a minimum. As changes in κ_g correspond to stretching the mid-plane of the leaf¹⁴, these observations imply that closure is characterized by the slow storage of elastic energy followed by its rapid release.

To understand the origin of these curvature changes, we measured local strains by recording the position of fiducial markers over the entire outer surface of a leaf before and after closure (see Methods). Our measurements of the strain (Fig. 2a) are consistent with earlier point-wise measurements^{5,12}, but go beyond these by characterizing the spatial structure of the strain field over the entire leaf. Figure 2a shows that the maximum strain perpendicular to the midrib (x-direction) is six times the maximum strain parallel to

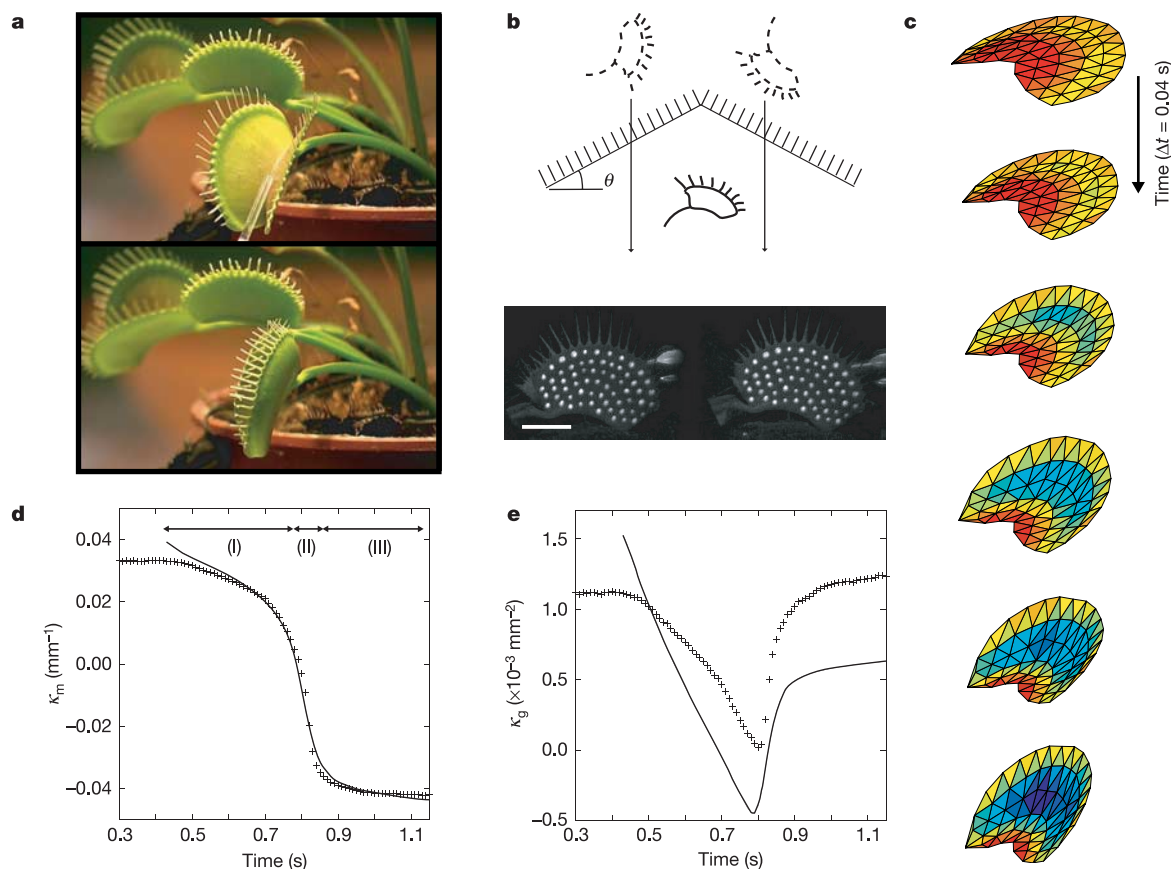


Figure 1 Dynamics of Venus flytrap closure. **a**, The Venus flytrap in its open and closed states. **b**, Schematic diagram of the imaging technique and typical stereo image showing fluorescent dots on the leaf surface under ultraviolet light. **c**, Dynamic sequence of the leaf closure. The time between images is 0.04 s. Colour indicates the value of the local mean curvature (blue $\kappa_m < 0$, red $\kappa_m > 0$). **d**, **e**, The spatially averaged mean curvature κ_m (**d**)

and the spatially averaged gaussian curvature κ_g (**e**) as a function of time. The plant was triggered at $t = 0$. The closure dynamics are characterized by three phases (I–III): a slow initial phase, a rapid intermediate phase and finally a second slow phase. The solid line corresponds to the theoretical model (see Methods). Scale bar, 1 cm in (**a**) and (**b**).

the midrib (y -direction). Furthermore, we find that the strains on the inner surface of the leaf are $\leq 1\%$, implying that closure is triggered primarily by differential strains in the x -direction. To corroborate this picture, we cut recently-closed leaves to determine the residual strains in them. Cutting a closed leaf in the x -direction eliminates the constraining effect of curvature in the y -direction and allows it to recover its natural curvature in the x -direction, κ_{xn} , as seen in Fig. 2b. Similarly, cutting the leaf in the y -direction allows us to observe the natural curvature in the y -direction, κ_{yn} . We see that κ_{xn} reverses sign during closure, but κ_{yn} does not. This is consistent with previous observations⁵ and provides quantitative evidence that a change in κ_{xn} drives leaf closure. Evidence supporting a mechanistic basis for this anisotropic deformation comes from

microscopic examination of the leaf surface (see Supplementary Fig. 1); we observe that the cells on the outer surface are highly elongated in shape, with their long axis oriented along the x -axis. As the cylindrical cell wall is reinforced azimuthally by microfibrils⁵, it follows that any changes in turgor would lead to deformations that are primarily in the x -direction, consistent with our macroscopic observations that κ_{xn} changes much more than κ_{yn} .

Although the molecular and cellular processes underlying the water movements that control anisotropic curvature changes remain poorly understood, we now argue that the macroscopic mechanism of closure is determined solely by leaf geometry. For a doubly-curved leaf (one that is curved in two orthogonal directions), bending and stretching modes of deformations are coupled¹⁴, meaning that bending the leaf causes its mid-plane to be stretched. If the coupling is weak, the leaf can change its shape from open to closed by varying its gaussian curvature and stretch without a large energetic cost. In such a situation, the leaf deforms smoothly to accommodate the change in κ_{xn} . If the coupling is strong, the leaf will not deform much (owing to the large energetic cost of stretching its mid-plane), until eventually the change in κ_{xn} becomes so large that the leaf snaps shut rapidly (Fig. 3). To quantify the smooth–snapping transition, we modelled the leaf as a thin, weakly-curved elastic shell with principal natural curvatures κ_{xn} (which changes) and κ_{yn} (which remains constant) and initial curvatures $\kappa_x(t=0) = \kappa_y(t=0) = \kappa$ (see Methods). The shape of

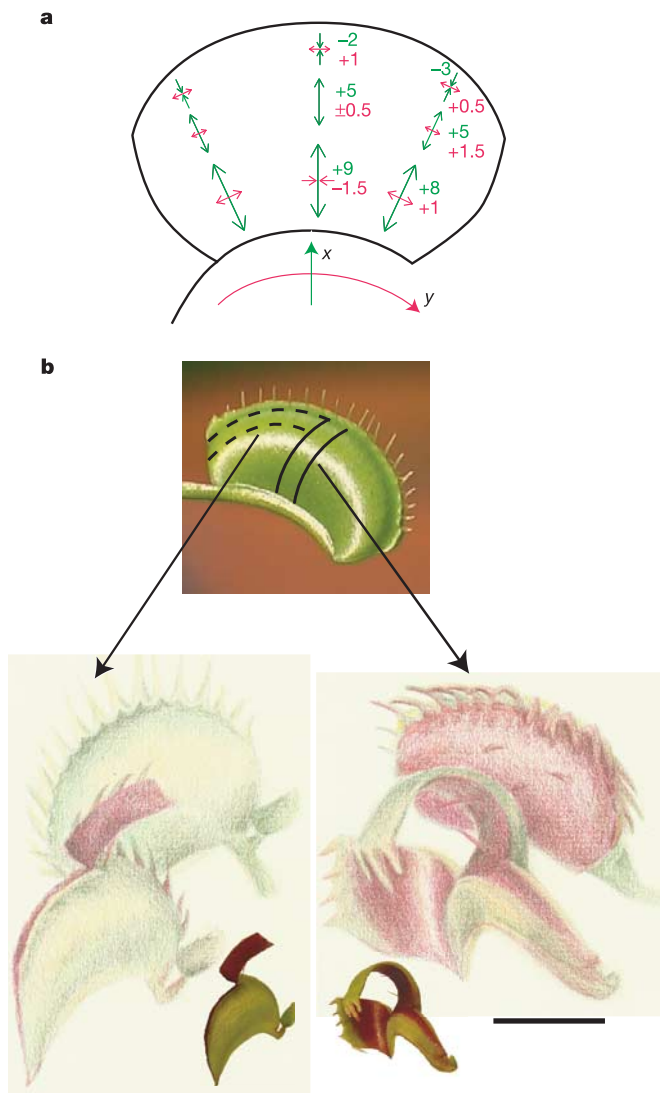


Figure 2 Strain field and natural curvature. **a**, Measured strain field of the outer face due to closure, where positive strains correspond to extension and negative strains correspond to contraction of the surface in the principal directions as shown. The maximum strain of 9% is perpendicular to the midrib (x -direction), whereas the strain parallel to the midrib (y -direction) is much smaller ($\pm 2\%$). **b**, Cutting the closed leaf along the dotted lines eliminates the coupling between bending and stretching in the doubly-curved leaf and shows κ_{xn} and κ_{yn} , the natural curvatures in the respective x - and y -directions. We find that κ_{yn} remains unchanged during leaf closure, but κ_{xn} changes actively and provides the motive force for snapping. The illustrations show the cut leaf overlaid on the uncut leaf; the smaller images show the cut leaf. For this leaf, the inner surface is red and the outer surface is green. Scale bar is for the illustration, 1 cm.

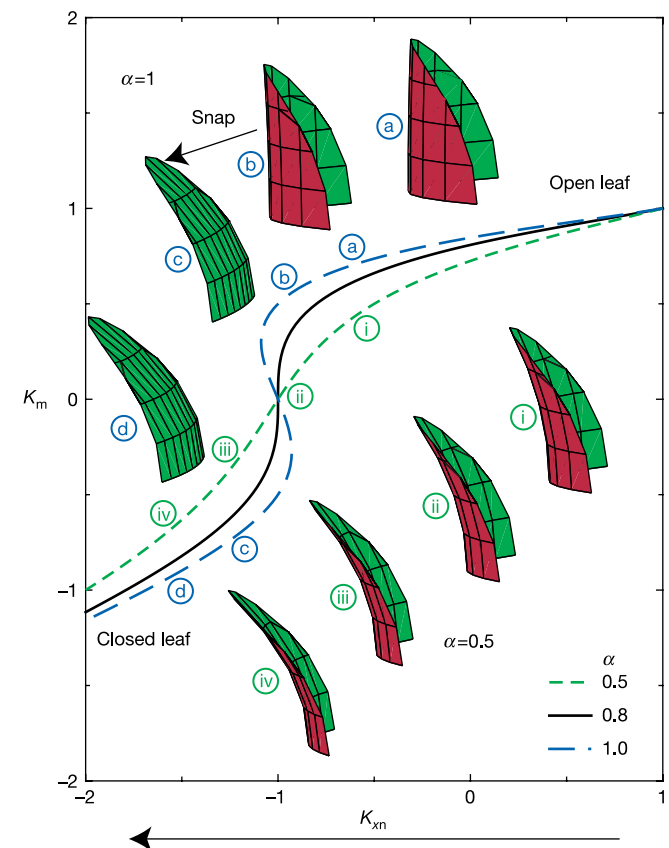


Figure 3 Smooth–snapping transition in leaf closure. We plot the dimensionless mean curvature $K_m = (K_x + K_y)/2$ as a function of the control parameter K_{xn} (see Methods). For $\alpha < \alpha_c = 0.8$ we observe a smooth transition from the open to the closed state, but for $\alpha > \alpha_c$ the system passes through a region of bistability before undergoing a rapid, ‘snapping’ transition. In the lower part of the figure we provide a schematic of the continuous transition ($\alpha = 0.5$); the upper part of the figure shows a similar schematic for the ‘snapping’ transition ($\alpha = 1$). Red denotes the inner surface and green denotes the outer surface of the leaf.

the leaf is found by minimizing the dimensionless elastic energy $U(\kappa_x, \kappa_y; \kappa_{xn}, \alpha) = U_{\text{bending}} + \alpha U_{\text{stretching}}$, which depends on the dimensionless natural curvature in the x -direction (κ_{xn} , which is controlled by the plant), and also depends on the dimensionless geometric parameter $\alpha = L^4 \kappa^2 / h^2$, which quantifies the coupling between bending and stretching deformations in terms of the leaf thickness h , the leaf size L , and the observed curvature of the open leaf κ (see Methods). A larger value of α implies that it is relatively more difficult to stretch the mid-plane by changing the gaussian curvature of the leaf. In Fig. 3 we see that $U(\kappa_x, \kappa_y; \kappa_{xn}, \alpha)$ has a single minimum when $\alpha \leq \alpha_c$ (weak bending–stretching coupling), but has two minima when $\alpha > \alpha_c$ (strong bending–stretching coupling) for some values of κ_{xn} . Our model thus predicts a region of bistability in which both the open and the closed states of the leaf are stable to small perturbations.

Consistent with this prediction, we have observed leaves that begin to close when stimulated, but do not snap or reach the closed state. However, by squeezing the leaves with our fingers we were able to induce a snap transition mechanically, indicating that the leaf was indeed in a bistable configuration. Because the amount of mid-plane stretching required for the leaf to close increases with α , the mechanical barrier to snapping is determined by this geometrical parameter. Thus we expect that large, highly curved leaves will release more energy and snap more rapidly than smaller, weakly curved leaves. To test this, we measured the speed of leaf closure as a function of the geometrical parameter α for many different leaves and plants. Figure 4a shows a clear positive correlation between the maximum speed of the snap and α . Interestingly, the time delay

between the beginning of the motion and the snap increases with the parameter α (Fig. 4b), as does the maximum speed of snapping. This is probably because the delay reflects the time required to cross the stretching barrier, which increases with α . For a typical leaf $\alpha = L^4 \kappa^2 / h^2 \approx (1 \text{ cm})^4 (0.1 \text{ cm}^{-1})^2 / (0.1 \text{ cm})^2 \approx 1 \approx \alpha_c$, suggesting that the leaves are poised at the transition between smooth closing and snapping. As the plant is clearly best served both by minimizing the delay between triggering and closure, and snapping rapidly rather than closing smoothly, this value of α might not be a coincidence.

The actual speed of snapping (the leaf closure time) is a critical parameter for trap function. Observations show that the motion is rapidly damped once the trap is closed. As the timescale of the fundamental bending mode of the leaf $\tau_B \approx (L^2 h^{-1})(\rho/E)^{1/2} \approx 0.001\text{--}0.01 \text{ s}$ is much less than the observed timescale of the motion ($\sim 0.1\text{--}1 \text{ s}$), and we observe no ringing, we conclude that inertial effects are not relevant in setting the timescale of motion. Furthermore, on comparing the elastic force acting on the leaf, F_E , to the force due to motion through the external air, F_A , we see that $F_A/F_E \approx \rho_f V^2 L^2 / Eh^3 \kappa \approx 10^{-6}$, where $V = 1 \text{ cm s}^{-1}$ and ρ_f is the density of air. Hence the damping must occur internally. Because the leaf is highly hydrated, a natural candidate for damping is provided by the flow of interstitial water through the surrounding elastic tissue. The curvature change during closure induces a transient flow perpendicular to the leaf surface, which occurs over a timescale $\tau_p \approx \mu h^2 / kE$ (refs 15, 16), where k is the hydraulic permeability of the leaf tissue, μ is the interstitial fluid viscosity and E is the drained bulk modulus of the tissue. Using a measured value of Young's modulus $E \approx 10 \text{ MPa}$ and taking $k/\mu \approx 10^{-12} \text{ m}^2 \text{ Pa}^{-1} \text{ s}^{-1}$ (ref. 17) we find that $\tau_p \approx 0.1 \text{ s}$, consistent with the snapping time. Our experiments on the response of the leaf tissue to impulsive and step loads are consistent with these estimates (see Supplementary Methods). A quantitative model that balances the change in elastic energy with the energy dissipated by fluid flow captures the three phases of snapping (Fig. 1d, e; see Methods and Supplementary Methods).

Our kinematic and mechanical measurements allow us to identify two distinct pieces in the puzzle of trap snapping: an active biochemical component and a passive elastic component. Upon stimulation, the plant 'actively' changes one of its principal natural curvatures, κ_{xn} , the microscopic mechanism for which remains poorly understood. Once this change occurs, the geometry of the doubly-curved leaf provides the mechanism by which elastic energy is both stored and released, and the hydrated nature of the leaf induces the rapid damping that is equally crucial for efficient prey capture. A single geometrical parameter (α) determines the nature of closure: if $\alpha \leq \alpha_c \approx 0.8$, the leaf closes smoothly, and if $\alpha > \alpha_c$, the leaf snaps rapidly. This ingenious solution to the problem of scaling up movements and speed from the cellular to the organ level in plants, nature's consummate hydraulic engineers, shows how controlling elastic instabilities in geometrically slender objects provides an alternative to the more common muscle-powered movements in animals. □

Methods

Stereoscopic reconstruction and curvature measurements

Plants were grown at room temperature (25 °C) and received 12 h of fluorescent light daily. All experiments were performed on healthy adult specimens. To record the three-dimensional (3D) shape of the leaf during closure, freshly excised traps were positioned in front of a pair of mirrors 2 m from a high-speed camera (Phantom V; Fig. 1b). Excised traps behaved as uncured traps as long as the time between excision and triggering was less than a few minutes. The stereo videos were post-processed using ImageJ software (<http://rsb.info.nih.gov/ij/>) to give the coordinates of the fluorescent dots for both left and right images. The corresponding 3D coordinates were obtained using a far field triangulation scheme (see Supplementary Methods for details).

We used two different approaches to compute local curvatures on the reconstructed leaf surface: first, global interpolation of the surface using p-splines (Matlab software) with the curvatures computed using finite differences, and second, local quadratic approximation of the surface¹⁸. Although the two methods gave similar qualitative

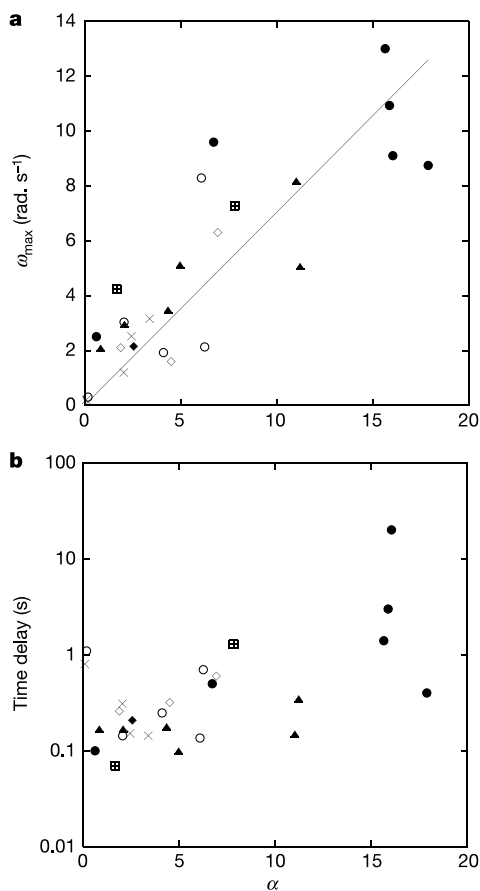


Figure 4 Speed of closure and delay between triggering and snapping. **a**, Maximum angular velocity ω plotted against α . **b**, The time between triggering the leaf and snapping, rapid closure plotted against α (note the semi-logarithmic scale). Identical symbols correspond to different leaves on the same plant.

outcomes, quantitative differences between the two results (~30%) arose due to intrinsic noise in the reconstruction and the limited number (~50) of points per leaf.

Strain field measurements

Moulds of the outer and inner surfaces of six different leaves were made before and after closure using a non-invasive replica technique¹⁹. We used a high-resolution digital camera fitted with a microscope lens to track the microscopic hairs of the outer face and the digestive glands of the inner face. The local strain field associated with closure (that is, the principal strains and the corresponding principal directions) was computed by local orientation-averaging of the strains (see Supplementary Methods) before determining their maxima and minima.

Measurement of α , maximum angular velocity and time delay

$\alpha = L^4 \kappa^2 / h^2$ was computed experimentally using $L = (L_1 + L_2)/4$, where L_1 is the length across the leaf measured parallel to the midrib and L_2 is the length measured perpendicular to the midrib. Measuring L/h for over 30 traps, we find that $L/h \approx 20$. κ is the initial mean curvature of the leaf. The maximum speed of the snap is the maximum angular speed of the vector perpendicular to the midrib that extends to the centre of the leaf's edge. The time delay is the time it takes the mean curvature to change from 5% to 50% of the total mean curvature change.

Elastic shell statics

We modelled the leaf as a thin, shallow shell with radius L , uniform thickness h , Young's modulus E , and natural curvatures κ_{xn} , κ_{yn} . Here we outline a minimal theory that captures the essential feature of these complex theories, namely that bending a doubly-curved shell leads to stretching because of changes in the gaussian curvature. This bend-stretch coupling is formalized in equations that describe the finite deformations of plates and shallow shells²⁰. Postponing the consideration of a complete theory, here we opt for a simplified energetic treatment that neglects the role of Poisson contraction and extension to enable us to focus on the essential mechanisms involved. Experimental data shows us that we may approximate the leaf shape as an elliptic (or hyperbolic if the two principal curvatures have opposite sign) paraboloid given by $f = (\kappa_x x^2/2) + (\kappa_y y^2/2)$, where κ_x and κ_y are the spatially homogeneous principal curvatures. Consistent with our experimental observations, we take the natural curvature in the y -direction to be a constant, equal to the initial curvature in the x - and y -directions before snapping, so that $\kappa_x(t=0) = \kappa_y(t=0) = \kappa_{ym}(t) = \kappa$. Then we can write the dimensionless total energy due to bending and stretching²⁰:

$$U(K_x, K_y; K_{xn}, \alpha) = U_{\text{bending}} + \alpha U_{\text{stretching}} = (K_x - K_{xn})^2 + (K_y - 1)^2 + \alpha(K_x K_y - 1)^2$$

where $K_x = \kappa_x/\kappa$, $K_y = \kappa_y/\kappa$ and $K_{xn} = \kappa_{xn}/\kappa$ are dimensionless. The geometric coupling parameter $\alpha = L^4 \kappa^2 / h^2$ characterizes the relative energy penalty of bending-to-stretching deformations. Minimizing $U(K_x, K_y; K_{xn}, \alpha)$ with respect to K_x , K_y ; that is, solving:

$$\frac{\partial U}{\partial K_x} = \frac{\partial U}{\partial K_y} = 0$$

for different values of α yields the change in leaf shape as a function of K_{xn} . When $\alpha > \alpha_c \approx 0.8$, there is just a single minimum, whereas for a typical leaf when $\alpha \approx 1$, we see the appearance of a second minimum (Fig. 3).

Poroelastic shell dynamics

To model the dynamics of snapping, we treat the leaf tissue as a poroelastic material. When the leaf snaps shut, the changes in curvature cause the movement of interstitial water relative to the elastic tissue, which dissipates energy. For an impermeable poroelastic plate with dimensionless curvatures $K_x(t)$ and $K_y(t)$ and fluid volume fraction ψ (ref. 16), balancing the elastic power due to curvature changes with the viscous dissipation rate due to fluid flow yields the dimensionless energy-balance equation (see Supplementary Methods):

$$\frac{dU}{dT} = A \left(\frac{\partial K_x(T)}{\partial T} + \frac{\partial K_y(T)}{\partial T} \right) \int_0^T e^{-(T-T')} \left(\frac{\partial K_x(T')}{\partial T'} + \frac{\partial K_y(T')}{\partial T'} \right) dT'$$

where $A = 196/\pi^4$. Here we have used a scaled time $T = t/\tau_p$, where $\tau_p = (\psi \mu h^2)/(\pi^2 k E)$ is the poroelastic time, and have scaled the curvatures and elastic energy (U) as in the previous section. We assume that K_x and K_y follow the path of steepest descent in the energy potential. In the absence of microscopic physiological information for the change in dimensionless natural curvature upon x -direction changes, we chose the form $K_{xn} = 1 - \alpha(1 - e^{-bT})$. To fit the data in Fig. 1d, we used $a = 3$, $b = 0.05$, $\tau_p = 1/94$ s and $\alpha = 1$. Other functional forms (for example, $K_{xn} = 1 - \alpha T$) do not change the qualitative dynamics of snapping, although quantitative aspects of the waiting time before the snap and the post-snap dynamics are affected.

Received 20 June; accepted 12 November 2004; doi:10.1038/nature03185.

1. Darwin, C. *Insectivorous Plants* (Murray, London, 1875).
2. Burdon-Sanderson, J. On the electromotive properties of the leaf of *dionaea* in the excited and unexcited states. *Phil. Trans. R. Soc. Lond.* **173**, 1–55 (1882).
3. Stuhlman, O. Jr & Darder, E. B. The action potentials obtained from Venus's flytrap. *Science* **111**, 491–492 (1950).
4. Hodick, D. & Sievers, A. The action potential of *Dionaea muscipula* Ellis. *Planta* **174**, 8–18 (1988).
5. Hodick, D. & Sievers, A. On the mechanism of trap closure of Venus flytrap (*Dionaea muscipula* Ellis). *Planta* **179**, 32–42 (1989).
6. Sibbald, T. Physiology of rapid movements in higher plants. *Annu. Rev. Plant Physiol.* **20**, 165–184 (1969).
7. Juniper, B. E., Robins, R. J. & Joel, D. M. *The Carnivorous Plants* (Academic, London, 1989).
8. Williams, S. E. & Bennet, A. B. Leaf closure in the Venus flytrap: an acid growth response. *Science* **218**, 1120–1122 (1982).

9. Stuhlman, O. A physical analysis of the opening and closing movements of the lobes of Venus' flytrap. *Bull. Torrey Bot. Club* **7**, 22–44 (1948).
10. Ashida, J. Studies on leaf movement of *Aldrovanda vesiculosa*. *Mem. Coll. Sci. Kyoto Imp. Univ., Ser. B* **9**, 141–244 (1934).
11. Morillon, R., Liénard, D., Chrispeels, M. J. & Lassalles, J.-P. Rapid movements of plants organs require solute-water cotransporters or contractile proteins. *Plant Physiol.* **127**, 720–723 (2001).
12. Brown, W. H. The mechanism of movement and duration of the effect of stimulation in the leaves of *dionaea*. *Am. J. Bot.* **3**, 68–90 (1916).
13. Hill, B. S. & Findlay, G. P. The power of movement in plants: the role of osmotic machines. *Q. Rev. Biophys.* **14**, 173–222 (1981).
14. Love, A. E. H. *A Treatise on the Mathematical Theory of Elasticity* 4th edn (Dover, New York, 1944).
15. Biot, M. A. General theory of three-dimensional consolidation. *J. Appl. Phys.* **12**, 155–165 (1941).
16. Skotheim, J. M. & Mahadevan, L. Dynamics of poroelastic filaments. *Proc. R. Soc. Lond. A* **460**, 1995–2020 (2004).
17. Frensch, J. & Steudle, E. Axial and radial hydraulic resistance to roots of maize (*Zea mays* L.). *Plant Physiol.* **91**, 719–726 (1989).
18. Dumais, J. & Kwiatkowska, D. Analysis of surface growth in shoot apices. *Plant J.* **31**, 229–241 (2002).
19. Green, P. B., Havelange, A. & Bernier, G. Floral morphogenesis in *Anagallis*: Scanning-electron-micrograph sequences from individual growing meristems before, during, and after the transition to flowering. *Planta* **185**, 502–512 (1991).
20. Mansfield, E. H. *The Bending and Stretching of Plates* 2nd edn (Cambridge Univ. Press, Cambridge, 1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. Shindler for the illustrations in Fig. 2. We acknowledge support via the Norwegian Research Council (J.M.S.) and the Schlumberger Chair Fund (L.M.) at Cambridge University, where this work was begun and primarily done.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.M. (lm@deas.harvard.edu).

Simulation and validation of modelled sphingolipid metabolism in *Saccharomyces cerevisiae*

Fernando Alvarez-Vasquez^{1,2}, Kellie J. Sims^{1,2}, L. Ashley Cowart¹, Yasuo Okamoto¹, Eberhard O. Voit^{1,2,3} & Yusuf A. Hannun¹

¹Department of Biochemistry and Molecular Biology and ²Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, South Carolina 29425, USA

³The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory Medical School, 313 Ferst Drive, Atlanta, Georgia 30332-0535, USA

Mathematical models have become a necessary tool for organizing the rapidly increasing amounts of large-scale data on biochemical pathways and for advanced evaluation of their structure and regulation. Most of these models have addressed specific pathways using either stoichiometric¹ or flux-balance analysis², or fully kinetic Michaelis-Menten representations³, metabolic control analysis⁴, or biochemical systems theory^{5–7}. So far, the predictions of kinetic models have rarely been tested using direct experimentation. Here, we validate experimentally a biochemical systems theoretical model of sphingolipid metabolism in yeast⁸. Simulations of metabolic fluxes, enzyme deletion and the effects of inositol (a key regulator of phospholipid metabolism) led to predictions that show significant concordance with experimental results generated post hoc. The model also allowed the simulation of the effects of acute perturbations in fatty-acid precursors of sphingolipids, a situation that is not amenable to direct experimentation. The results demonstrate that modelling now allows testable predictions as well as the design and evaluation

of hypothetical 'thought experiments' that may generate new metabolomic approaches.

Recent recognition of sphingolipids, such as sphingosine, ceramide and sphingosine-1-phosphate, as bioactive molecules^{9–11} has increased interest in understanding sphingolipid metabolism and its regulation. Such understanding is impeded, however, by the complexity and multitude of interconnected reactions that characterize the sphingolipid pathway¹², thus highlighting the need for mathematical modelling⁸. We selected yeast as our target organism for modelling because sphingolipid metabolism is largely conserved among eukaryotes. Moreover, all primary metabolites, enzymes and their genes, and some of the regulatory signals, are known in this

organism¹³, thus allowing the crucial initial step of model design through the proper mapping of the metabolic pathways as shown in Fig. 1 (and listed in detail in Supplementary Tables S1 and S2). Additionally, the availability of mutant yeast strains for many genes of sphingolipid enzymes facilitates experimental verification of model predictions. The details of model design, including the rationale for using biochemical systems theory (BST), incorporation of complex sphingolipids, and fatty-acid synthesis and elongation, are presented and discussed in the Supplementary Text 1 and 2.

Once the model was assembled and proofed, it was critical to probe for parameters that could unduly influence the model. This

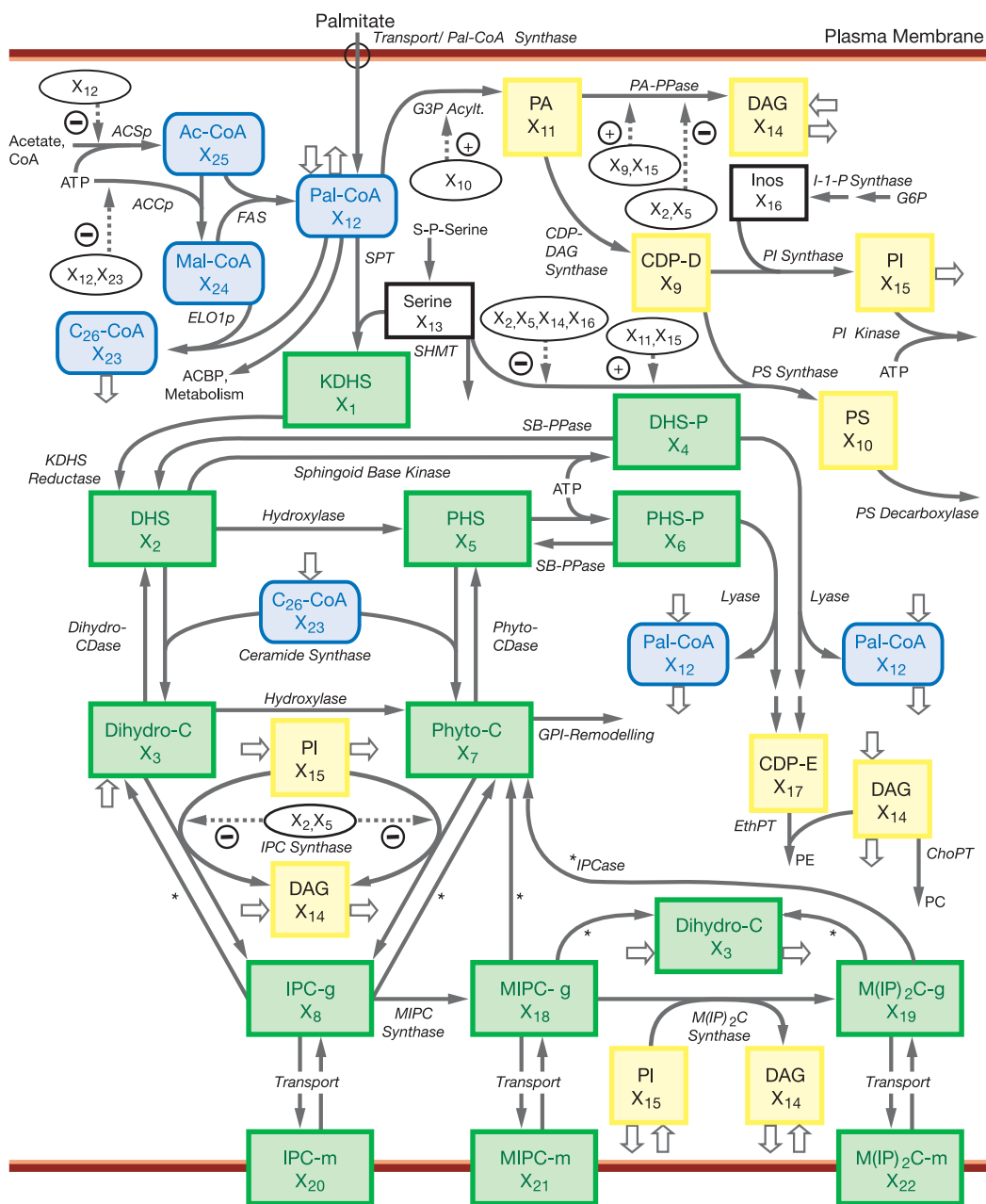


Figure 1 Metabolic map of sphingolipid metabolism and related processes in yeast. Dependent variables (metabolites and cofactors) are boxed and independent variables (enzymes and constant components) are in italics. The corresponding index notation for each variable in the GMA system is also given. Arrows indicate material fluxes (solid) and signals (dashed). Open arrows indicate a connection to another area of the map. Putative

regulatory signals are presented in ellipses. Sphingolipid metabolism is shown in green, fatty-acid metabolism in blue, and phospholipid metabolism in yellow. The asterisk shows the multiple locations for Isc1, the inositol sphingolipid phospholipase C (IPCase). Definition of abbreviations and gene products are listed in Supplementary Tables 1 and 2.

was achieved by performing sensitivity analysis, and the results showed that the system is robust at steady state. The vast majority of sensitivities are small in magnitude (Supplementary Table S3), indicating attenuation or modest amplification of most perturbations. The remaining high sensitivities can be explained and fall into two distinct groups (Supplementary Fig. 3). The first contains some of the central metabolites of sphingolipid metabolism, indicating a capacity for responsiveness that is expected in signal transduction systems. The second group possesses the highest sensitivities and comprises metabolites on the periphery of the pathway where the model representation is coarser and more susceptible to perturbations. Details can be found in Supplementary Text 3. Overall, the extended model was deemed ready for testable predictions and validation.

The validation experiments typically tracked sphingolipid fluxes, using radiolabelled precursors. To allow comparisons of experimental and simulated results, we quantified the loss of radioactive tracer from the medium over the time course of the experiment. Specifically, the loss was fitted with a standard decay function and used in the model as cellular substrate uptake (see Supplementary Text 4). The flux of radioactive tracer through the sphingolipid pathway was simulated with a novel mathematical method¹⁴ specifically developed for this purpose (Supplementary Text 5).

The first simulation followed the dynamics of incorporation of [³H]-palmitate into sphingolipids (bulk concentration of palmitate of 0.182 μ M, specific activity of 53 Ci mmol⁻¹) over a time course of 60 min (Fig. 2a). The simulation showed an immediate increase in labelled dihydrosphingosine (DHS), which then gradually declined. Complex sphingolipids exhibited a delayed increase that was sustained throughout the time course.

Figure 2b shows the corresponding experimental results. Tritiated

palmitate (0.182 μ M) was added to the medium of mid-log-phase yeast. Samples were taken at various time points, lipids were extracted and separated by thin-layer chromatography (TLC), and individual bands were quantified by liquid scintillation (see Methods). As seen, DHS increased early on and then decreased at later time points, whereas the complex sphingolipids increased steadily over the entire time course, as predicted by the model.

Sphingolipids have only one known route of ultimate breakdown, namely through the lyase encoded by *DPL1* (ref. 15); this prompted an investigation of the effects of deleting this gene. The simulation with the BST model was performed by eliminating the appropriate flux term to mimic the mutant, and then following the palmitate tracer in the wild type and the *DPL1* knockout. The model predicted the levels of long-chain base phosphates (LCB-P)—the substrates of the lyase—to be significantly higher in the knockout than in the wild type at the early time points, but eventually to return to wild-type levels by clearance through the action of the LCB-P phosphatases (Fig. 2c). The levels of DHS and complex sphingolipids in wild type and knockout were predicted by the model to be similar (Fig. 2c, e).

In the corresponding experiment, wild type and *DPL1* knockout strains were labelled with [³H]-palmitate, and the radioactivity was measured in individual sphingolipid species. The levels of the LCB-Ps were appreciably elevated in the knockout but not the wild type, where they were below detection limits (Fig. 2d). This agrees with the miniscule amounts of dihydrosphingosine-1-phosphate (DHS-1-P) and phytosphingosine-1-phosphate (PHS-1-P) (0.0046% of total phospholipids) previously reported¹⁶. The autoradiographs (Supplementary Fig. 5) show prominent bands for the LCB-Ps at the 5-min time point in the knockout, but not in the wild type. These bands decreased within 40 min, as in the simulation. Levels of DHS in the wild type and lyase knockout followed a similar shape, but the wild type had slightly lower levels at each time point.

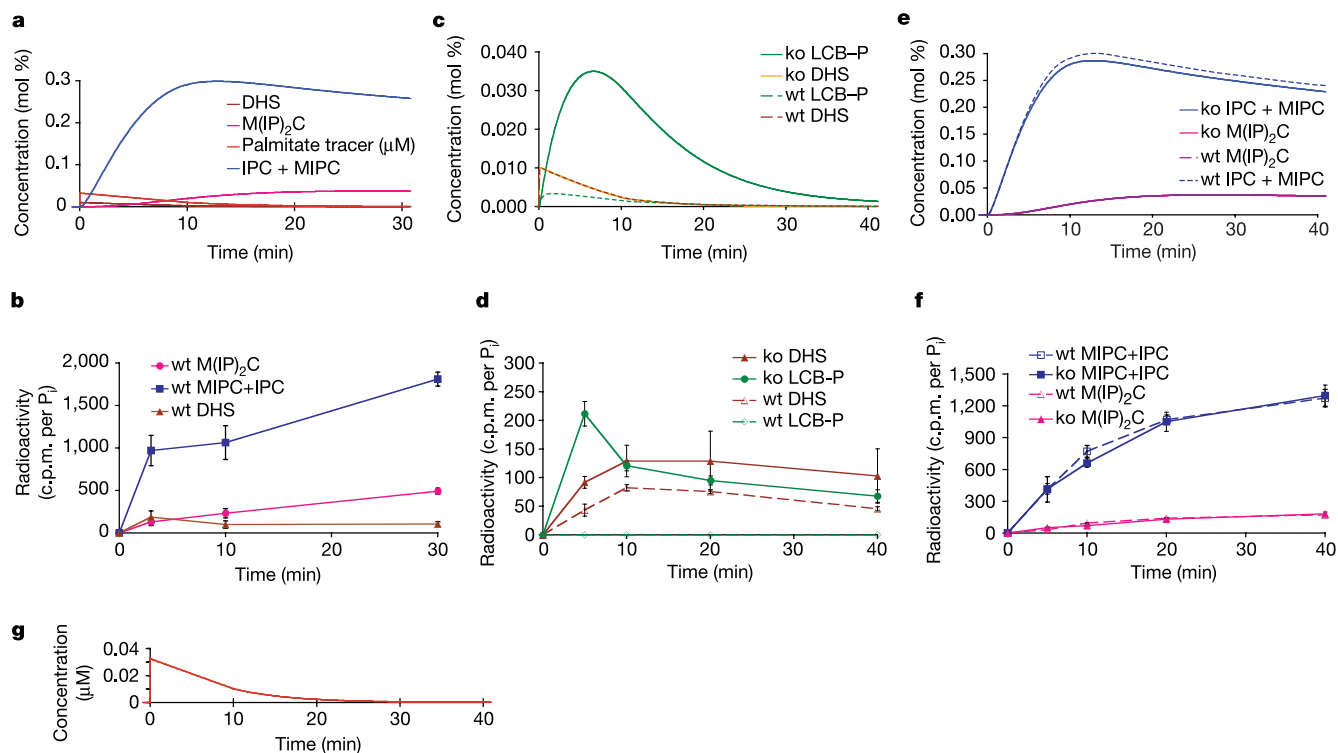


Figure 2 Experimental and simulated lipid profiles from labelled palmitate. **a**, Simulation of incorporation of [³H]-palmitate into various sphingolipids. MIPC; mannose-inositol-phospho-ceramide; M(IP)₂C, mannose-bis(inositol-phospho)-ceramide. **b**, Corresponding experimental labelling results. **c**, Simulated profile of DHS and LCB-P for wild type and lyase deletion mutants. **d**, Experimental profile of DHS and LCB-P for wild type and lyase

deletion mutants. **e**, Simulated profile of complex sphingolipids for wild type and lyase deletion mutants. **f**, Experimental profile of complex sphingolipids for wild type and lyase deletion mutants. **g**, Palmitate tracer in simulations **c** and **e**. Standard error bars in **b**, **d**, **f** represent two to four separate measurements.

The levels of complex sphingolipids were very close for wild type and knockout strains, and increased steadily in the first 20 min, before reaching a plateau (Fig. 2f and Supplementary Fig. 5). Thus, the dynamic model predictions were accurate. Furthermore, steady-state analysis of the model revealed that the basal PHS level in the lyase knockout should be ~20% greater than in wild type (data not shown), which fit well with reported experimental data¹⁷.

As a more complex application, we tested the responses of sphingolipids to large perturbations in overall lipid metabolism by simulating the addition of inositol, a well-studied and important regulator of (glycerol-) lipid metabolism in yeast¹⁸. Loewen *et al.* showed recently that addition of 75 μ M inositol markedly increased phosphatidylinositol (PI) and slightly lowered most other phospholipids, including phosphatidic acid, which was proposed to act as a bioactive lipid in regulating transcription responses to inositol. We speculated that the increased PI should increase synthesis of complex sphingolipids through *IPC1*, which utilizes PI as a substrate and hence consumes ceramide and reduces its levels. We simulated the effects of the external inositol by increasing the basal mass of PI in the model three times and followed the changes in ceramides and complex sphingolipids. The result was a significant increase in complex sphingolipids accompanied by a marked decrease in ceramide levels (Fig. 3a). Interestingly, we found a biphasic response for ceramide with an initial decrease followed

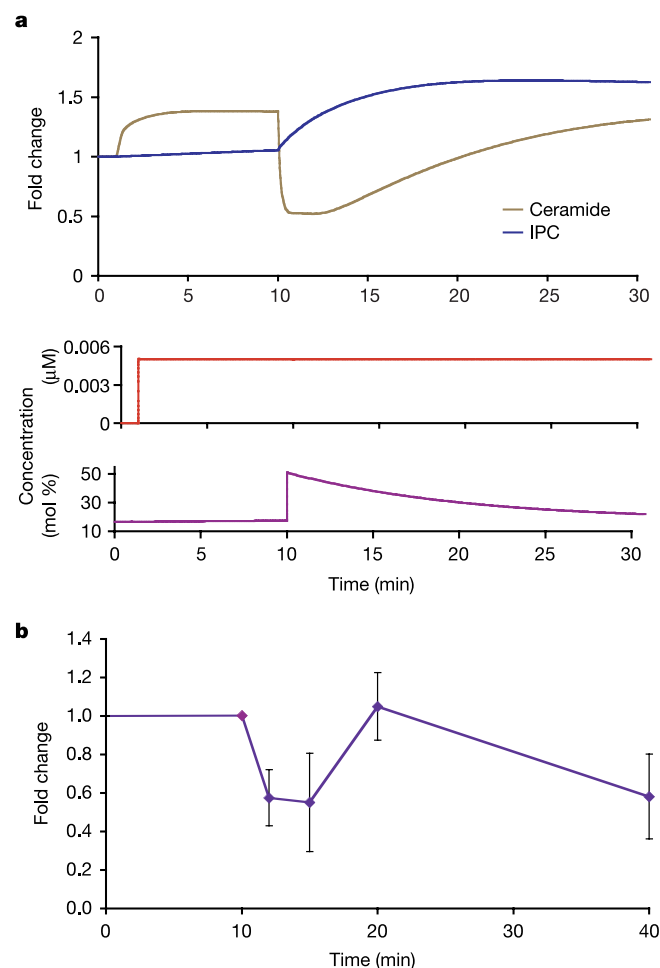


Figure 3 Response of sphingolipid pathway to increased inositol. **a**, Simulated profile after a threefold increase in the basal mass of PI at 10 min. The bottom panels show palmitate (red line) and phosphatidylinositol (purple line). **b**, Experimental profile of total ceramide levels after addition of 75 μ M inositol. Standard error bars represent two experiments.

by a return to baseline. To test the veracity of this simulation, we measured ceramide levels by mass spectrometry under similar conditions as reported¹⁸ and, indeed, found the same biphasic behaviour (Fig. 3b). Although it is not proof of correctness, this non-intuitive result suggests that the model functions adequately even for rather tangential aspects of sphingolipid metabolism. Also, because ceramide is a bioactive lipid, the results raise the intriguing hypothesis that acute and significant changes in ceramide may contribute to the cellular response to inositol.

The concordance between model and experimental results suggested the possibility of predicting the outcome of experiments that are otherwise difficult to perform or control. As an example, we simulated the effects of changes in endogenous palmitoyl-CoA on the flux through the *de novo* pathway. This part of the pathway is complicated because changes in the levels of palmitoyl-CoA may result from changes in either external palmitate or changes in the rate of synthesis of endogenous palmitate through changes in malonyl-CoA. A square pulse was used to increase the total concentration of each precursor to 1.5 times the basal mass level for one minute, followed by an immediate return to the initial value.

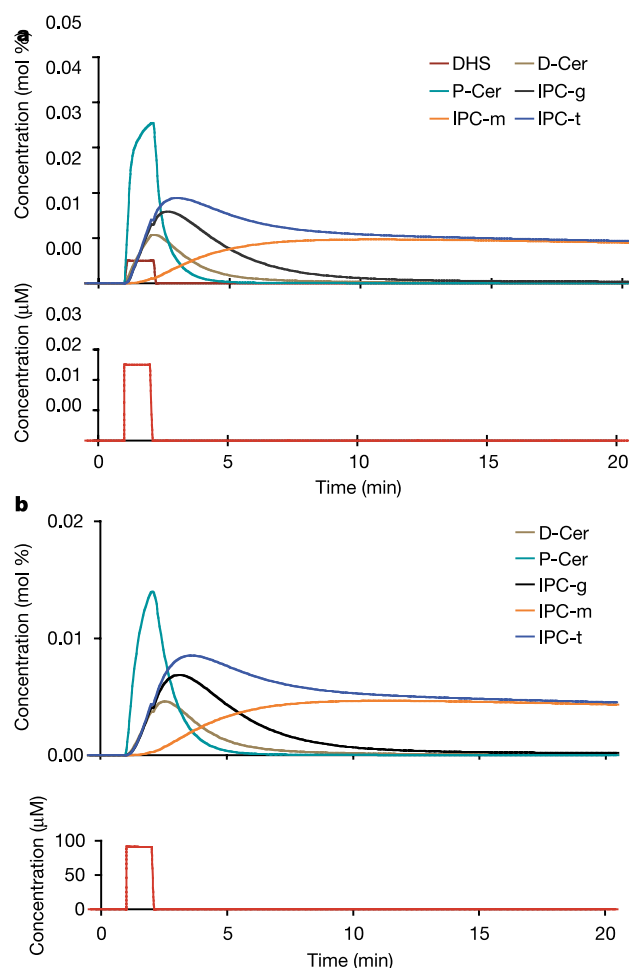


Figure 4 Response of sphingolipid pathway to perturbation of palmitoyl-CoA precursors. **a**, Simulated time course following bolus of external palmitate. **b**, Simulated time course following bolus of endogenous malonyl-CoA. A square pulse of tracer was simulated for palmitate and for malonyl-CoA. For each precursor, the value was initially set at baseline, raised to one-half the basal value of the bulk mass at steady state for one minute, and then returned to basal. The effect of this 'pulse-chase' input was followed with a dynamic GMA system of tracer equations¹⁴ (see Supplementary Text 5). D-Cer, *D-erythro*-dihydroceramide; P-Cer, *D-erythro*-phytoceramide; IPC-m/g/t, IPC at plasma membrane (m), Golgi (g) or total (t).

Each precursor value was perturbed separately (Supplementary Text 5), and the resulting lipid profiles were compared. The simulations showed that after changing the levels of either palmitate (Fig. 4a) or malonyl-CoA (Fig. 4b), the tracer quickly dissipated through the upstream metabolites (the sphingoid bases and ceramides) of the metabolic pathway whereas it accumulated and decreased more slowly in the downstream metabolites and, in particular, the complex sphingolipids.

One important difference between the perturbations of external (exogenous) palmitate versus internal (malonyl CoA) palmitate was the maximum amplitude of palmitoyl-CoA produced by each, as well as the resulting flux through the sphingoid backbones. The external palmitate perturbation produced an increase in palmitoyl-CoA of $4.9 \times 10^{-3} \mu\text{M}$ compared with $2.94 \times 10^{-8} \mu\text{M}$ from the malonyl-CoA (data not shown). This suggests a greater influence of external palmitate, as opposed to endogenous palmitoyl-CoA, on *de novo* sphingolipid synthesis. The fluxes of dihydroceramide and phytoceramide exhibited a similar shape in both simulated perturbations, but were reduced in amplitude by about two-thirds for the malonyl-CoA precursor (Fig. 4a versus 4b).

Taken together, these results suggest that the availability of the acyl-CoA substrate is very important for the first enzymatic reactions of *de novo* sphingolipid synthesis. This significant influence on sphingolipid production by the availability of acyl-CoA substrate can be explained by the fact that the free *in vivo* levels of long-chain acyl-CoA esters in yeast are maintained at nanomolar concentrations by the acyl binding protein (Acb1p), the sterol carrier protein 2 and the fatty-acid binding protein¹⁹. These three principal reservoirs of acyl-CoAs maintain a low free concentration of acyl-CoA in rapid equilibrium with the large pool of bound acyl-CoA. This equilibrium, in turn, allows large amounts of palmitoyl-CoA to be readily available for specific processes such as β -oxidation, glycerolipid synthesis and protein acylation, but without the need to maintain high concentrations of free acyl-CoA, which can be toxic and can inhibit key enzymes such as Acc1p (refs 20, 21).

The results from this study support the three main goals of biochemical pathway modelling. First, the current model provides a comprehensive, functional integration of experimental data (Supplementary Tables S1 and S2) that facilitates the understanding of the organization and regulation of sphingolipid metabolism in yeast. Second, the simulations show significant agreement with experimental results, thus establishing the utility of the model. Third, the model may be used to predict results of hypothetical experiments that are otherwise difficult, or even impossible, to perform but that yield insights that are critical for understanding the functionality of the system. Taken together, the model may serve as a rigorous foundation for a larger assemblage of information on yeast lipid metabolism and as a nidus for meaningful integration of metabolomic data. □

Methods

Yeast strains and culture media

The *S. cerevisiae* strains used were JK9-3d α (*MAT α trp1 leu2-3 his4 ura3 ade2rme1*) (ref. 22) for wild type and SGP3 (*MAT α leu2-3,112 trp1 ura3-52 his3 ade8 ras1::HIS3*) for wild type of the lyase deletion strain JS16 (*MAT α leu2-3,112 trp1 ura3-52 his3 ade8 ras1::HIS3 bst1 Δ ::NEO*) (ref. 9). Media used were YPD with 2% w/v glucose and synthetic minimal complete (SC). A single colony from a YPD plate of each strain was inoculated into 5 ml of YPD and incubated at 30 °C. Fresh YPD was used to dilute 24-h cultures to an absorbance of 0.8 at 600 nm ($A_{600} = 0.8$) and this was used to inoculate an appropriate amount of medium for the labelling experiment (1 ml medium per sample point + 1–2 ml extra) for overnight growth to mid-log phase in a shaker water bath at 30 °C and 220 r.p.m.

Sphingolipid labelling

Cells were resuspended in SC medium and allowed to equilibrate for 30 min in a shaker water bath at 30 °C and 220 r.p.m. before labelling with $1 \mu\text{Ci ml}^{-1}$ of $[9,10\text{-}^3\text{H}]$ -palmitate (53 Ci mmol^{-1}). Samples of 1 ml were taken at desired time points and immediately placed on ice with 100 μl 10% tricarboxylic acid (TCA).

Lipid extraction

Lipids were extracted as previously described²³.

Lipid separation and quantification

Lipid samples were resuspended in 50 μl of 16:16:5 chloroform:methanol:water and 25 μl was applied to a Whatman silica gel TLC plate, which was pre-treated with acetone. The plate was resolved using 9:7:2 chloroform:methanol:4.2 M ammonium hydroxide, sprayed with EN³HANCE and radiographed on BioMax film. Individual lipids were quantified by scraping the band and counting with liquid scintillation. Lipid measurements from TLC were normalized by total lipid phosphate, which was determined with a standard curve analysis and colorimetric assay of ashed phosphate as previously described²⁴.

Liquid chromatography/mass spectrometry analysis of yeast sphingolipids

Cells were cultured to mid-log phase in SC media. Myo-inositol was added to a final concentration of 75 mM, and samples were taken at the indicated times and collected by centrifugation at 2,800g. Cell pellets were snap-frozen in a methanol/dry ice bath and stored at $-80\text{ }^{\circ}\text{C}$. Internal standards were added to frozen pellets and sphingolipids were extracted in a one-phase neutral organic solvent (isopropanol:water:ethyl acetate 30:10:60 v:v:v) as described²⁵. A fraction of total extract from each sample was reserved for phosphate determination, and the remaining samples were then analysed by a Surveyor/TSQ 7000 liquid chromatography/mass spectrometry system. Lipids were qualitatively defined by parent ion scanning for known fragments characteristic for a specific sphingolipid class including sphingoid bases, sphingoid base phosphates and ceramides, as described²⁵. Samples were quantitatively analysed on the basis of calibration curves generated with synthetic standards. The mass of each species was normalized by organic phosphate determination.

Modelling

Software used for model development included Maple 7 for the computation of kinetic parameters and development of the generalized mass action (GMA) equations, and PLAS (ref. 26) for the numerical integration and analysis of the differential equations.

Received 22 June; accepted 26 November 2004; doi:10.1038/nature03232.

- Heinrich, R. & Schuster, S. *The Regulation of Cellular Systems* (Chapman and Hall, New York, 1996).
- Schilling, C. H., Edwards, J. S., Letscher, D. & Palsson, B. O. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.* **71**, 286–306 (2000).
- Mulquin, P. J. & Kuchel, P. W. *Modeling Metabolism with Mathematica: Analysis of Human Erythrocyte Metabolism* (CRC Press, Boca Raton, Florida, 2003).
- Fell, D. A. *Understanding the Control of Metabolism* (Portland, London, 1997).
- Savageau, M. A. Biochemical systems analysis. *J. Theor. Biol.* **25**, 365–379 (1969).
- Savageau, M. A. *Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology* (Addison-Wesley, Reading, Massachusetts, 1976).
- Voit, E. O. *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists* (Cambridge Univ. Press, Cambridge, 2000).
- Alvarez-Vasquez, F., Sims, K. J., Hannun, Y. A. & Voit, E. O. Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models. *J. Theor. Biol.* **226**, 265–291 (2004).
- Hannun, Y. Functions of ceramide in coordinating cellular responses in stress. *Science* **274**, 1855–1859 (1996).
- Hannun, Y. A. & Obeid, L. M. The ceramide-centric universe of lipid-mediated cell regulation: stress encounters of the lipid kind. *J. Biol. Chem.* **277**, 25847–25850 (2002).
- Spiegel, S. & Milsten, S. Sphingosine-1-phosphate: an enigmatic signalling lipid. *Nature Rev. Mol. Cell Biol.* **4**, 397–407 (2003).
- Hannun, Y. A., Luberto, C. & Argraves, K. M. Enzymes of sphingolipid metabolism: from modular to integrative signaling. *Biochemistry* **40**, 4893–4903 (2001).
- Sims, K. J., Spassieva, S. D., Voit, E. O. & Obeid, L. M. Yeast sphingolipid metabolism: clues and connections. *Biochem. Cell Biol.* **82**, 45–61 (2004).
- Voit, E. O., Alvarez-Vasquez, F. & Sims, K. J. Analysis of dynamic labeling data. *Math. Biosci.* **191**, 83–99 (2004).
- Saba, J. D., Nara, F., Bielawska, A., Garrett, S. & Hannun, Y. A. The *BST1* gene of *Saccharomyces cerevisiae* is the sphingosine-1-phosphate lyase. *J. Biol. Chem.* **272**, 26087–26090 (1997).
- Skrzypek, M. S., Nagiec, M. M., Lester, R. L. & Dickson, R. C. Analysis of phosphorylated sphingolipid long-chain bases reveals potential roles in heat stress and growth control in *Saccharomyces*. *J. Bacteriol.* **181**, 1134–1140 (1999).
- Zhang, X., Skrzypek, M. S., Lester, R. L. & Dickson, R. C. Elevation of endogenous sphingolipid long-chain base phosphates kills *Saccharomyces cerevisiae* cells. *Curr. Genet.* **40**, 221–233 (2001).
- Loewen, C. J. et al. Phospholipid metabolism regulated by a transcription factor sensing phosphatidic acid. *Science* **304**, 1644–1647 (2004).
- Knudsen, J. et al. Role of acyl-CoA binding protein in acyl-CoA transport, metabolism and cell signaling. *Mol. Cell. Biochem.* **192**, 95–103 (1999).
- Shirra, M. et al. Inhibition of acetyl coenzyme A carboxylase activity restores expression of the *INO1* gene in a snf1 mutant strain of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **21**, 5710–5722 (2001).
- Rasmussen, J., Rosendal, J. & Knudsen, J. Interaction of acyl-CoA binding protein (ACBP) on processes for which acyl-CoA is a substrate, product or inhibitor. *Biochem. J.* **292**, 907–913 (1993).
- Mao, C., Wadleigh, M., Jenkins, G. M., Hannun, Y. A. & Obeid, L. M. Identification and characterization of *Saccharomyces cerevisiae* dihydrospingosine-1-phosphate phosphatase. *J. Biol. Chem.* **272**, 28690–28694 (1997).
- Mandal, S. M. et al. Inhibition of serine palmitoyl-transferase activity by lipoxamycin. *J. Antibiot. (Tokyo)* **47**, 376–379 (1994).
- Jenkins, G. M. & Hannun, Y. A. in *Apoptosis: A Practical Guide* (ed. Studzinski, G. P.) (Oxford Univ. Press, New York, 1999).
- Bielawska, J., Szulc, Z. M., Hannun, Y. A. & Bielawska, A. *Simultaneous Quantitative Analysis of Bioactive Sphingolipids by High Performance Liquid Chromatography-Tandem Mass Spectroscopy in Methods* (ed. Bankaitis, V.) (Academic press, San Diego, in the press).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by a NIH grant. Doctoral studies of K.J.S. are supported by a NLM training grant. The authors would like to thank C. Mao and S. Vaena de Avalos for helpful discussions and the MUSC Lipidomics Core for sample analysis.

Authors' contributions F.A.-V. developed the model and conducted the simulations. K.J.S. performed the experiments under the supervision of Y.O. and drafted the manuscript. L.A.C. performed the inositol experiment. Y.A.H. and E.O.V. conceived and supervised the collaboration and overall strategy of the project, and edited the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.O.V. (eberhard.voit@bme.gatech.edu) or Y.A.H. (hannun@musc.edu).

APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses

Cécile Esnault^{1,2}, Odile Heidmann¹, Frédéric Delebecque², Marie Dewannieux¹, David Ribet¹, Allan J. Hance³, Thierry Heidmann^{1*} & Olivier Schwartz^{2*}

¹Unité des Rétrovirus Endogènes et Éléments Rétroviraux des Eucaryotes Supérieurs, UMR8122 CNRS, Institut Gustave Roussy, Villejuif 94805, France

²Virus and Immunity Group, Department of Virology, Institut Pasteur, CNRS URA1930, Paris 75015, France

³INSERM U552, Hôpital Bichat-Claude Bernard, Paris 75018, France

* These authors contributed equally to this work

Endogenous retroviruses are multicopy retroelements accounting for nearly 10% of murine or human genomes^{1,2}. These retroelements spread into our ancestral genome millions of years ago and have acted as a driving force for genome evolution²⁻⁴. Endogenous retroviruses may also be deleterious for their host, and have been implicated in cancers and autoimmune diseases⁵. Most retroelements have lost replication competence because of the accumulation of inactivating mutations, but several, including some murine intracisternal A-particle (IAP) and MusD sequences, are still mobile^{6,7}. These elements encode a reverse transcriptase activity and move by retrotransposition, an intracellular copy-and-paste process involving an RNA intermediate. The host has developed mechanisms to silence their expression, mainly cosuppression and gene methylation^{4,8}. Here we identify another level of antiviral control, mediated by APOBEC3G, a member of the cytidine deaminase family that was previously shown to block HIV replication⁹⁻¹². We show that APOBEC3G markedly inhibits retrotransposition of IAP and MusD elements, and induces G-to-A hypermutations in their DNA copies. APOBEC3G, by editing viral genetic material, provides an ancestral wide cellular defence against endogenous and exogenous invaders.

Mammalian long-terminal-repeat (LTR) retrotransposons (or endogenous retroviruses) such as murine IAP and MusD sequences or human endogenous retroviruses (HERVs) are structurally similar to infectious retroviruses^{1,2,4} (Fig. 1a). Their life cycle includes the formation of virus-like particles (VLPs), which, in several cases, remain intracellular¹. IAPs are assembled and bud at the ER membrane. IAPs are expressed in the germ line and in various tumour cells^{1,13,14}. MusD elements are transcribed at high levels in

undifferentiated embryonal carcinoma cells and in early embryos, but not in more differentiated tissues^{1,15}. Among the numerous IAP and MusD copies in the mouse genome, about 300 IAP and 10 MusD copies are thought to be still active, capable of retrotransposition^{6,7} (Fig. 1a). No active HERVs have yet been isolated, despite evidence for recent amplification in humans².

Retroelements share with infectious retroviruses a reverse transcription step. APOBEC3G, a member of the APOBEC family of nucleic-acid editing enzymes^{16,17}, is a powerful inhibitor of HIV and murine leukaemia virus (MLV) replication, acting during reverse transcription. APOBEC3G is a cytoplasmic protein that is incorporated into virions and which deaminates cytosine residues to uracils (C to U) in the nascent DNA. Most uracil-containing viral DNAs are then degraded, probably by cellular enzymes, before

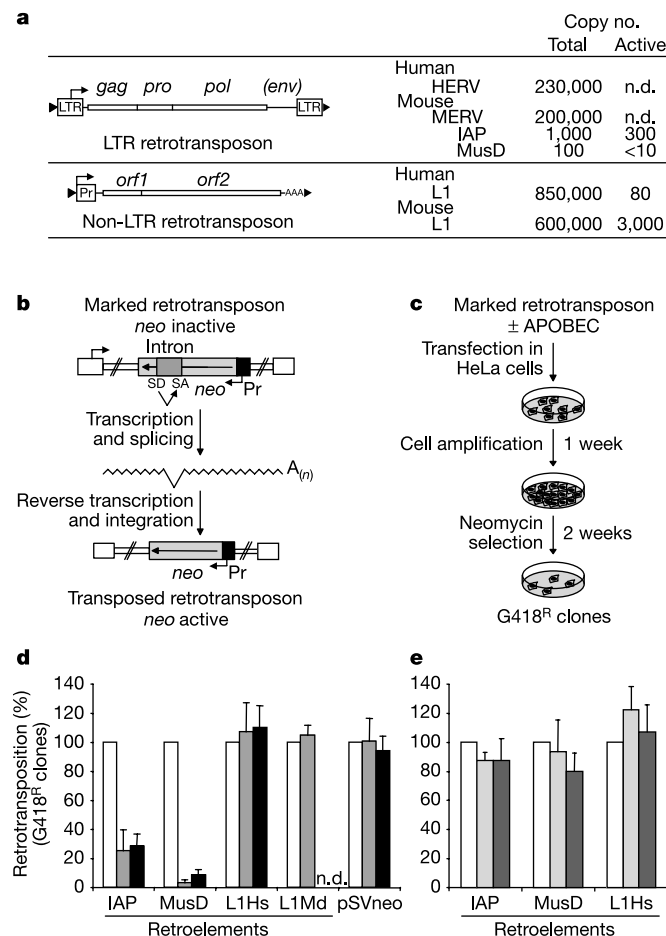


Figure 1 APOBEC3G inhibits the retrotransposition of endogenous retroviruses. **a**, The structure of retroelements and their estimated occurrence in murine and human genomes. LTR retrotransposons contain *gag*, *pro* and *pol* genes, and usually lack a functional *env* gene. L1 elements possess two open reading frames, *orf1* (encoding a nucleic-acid-binding chaperone protein) and *orf2* (encoding reverse transcriptase and endonuclease). Black triangles indicate target site duplications; n.d., not determined. **b**, Representation of a retroelement marked with the *neo* reporter gene placed in reverse orientation with its own promoter (Pr). Besides the *neo* cassette, each marked retroelement carries its set of functional genes, allowing autonomous retrotransposition. **c**, Experimental procedure for the detection of retrotransposition. **d**, Analysis of the activity of human APOBEC3G (hA3G; grey bars) and murine APOBEC3 (mA3; black bars) on the indicated retroelements. Retrotranspositions (defined by quantifying G418^R clones) are presented as percentages relative to samples containing retroelement alone (white bars). pSVneo, control *neo* expression plasmid. **e**, Murine APOBEC1 (mA1; light grey bars) and APOBEC2 (mA2; dark grey bars) do not affect retrotransposition of the indicated retroelements. White bars, no APOBEC. Data are means ± s.d. for three independent experiments.

integration. Some molecules escape degradation, and the resulting proviruses contain numerous guanosine-to-adenosine (G-to-A) substitutions in the plus strand^{9–12,18,19}. APOBEC3G, as well as some other members of the APOBEC family, can also induce C-to-T mutations in the HIV genome, although at a low frequency^{20–22}. Interestingly, the HIV Vif protein neutralizes the APOBEC3G-mediated antiviral effect by promoting proteasomal degradation of the enzyme^{23,24}.

The physiological role of APOBEC3G remains unknown. The APOBEC3G gene has been under strong positive selection throughout the past 30 million years of primate evolution, and predates HIV-like viruses²⁵. APOBEC3G might represent an ancient form of antiviral defence in mammals. We examined whether a primary function of this protein is to restrict retrotransposition of endogenous retroviruses. For this purpose we used a cell-based retrotransposition assay, in which the retroelement is marked with a *neo* cassette placed in reverse orientation (Fig. 1b). The *neo* gene is rendered inactive by the presence of a forward intron, which is spliced out of the RNA intermediate, resulting in a functional gene after reverse transcription and integration^{6,7,13,26}. Two LTR retrotransposons (murine IAP and MusD) were tested, for which functional copies have been cloned and showed to be autonomous for retrotransposition^{6,7}. HeLa cells were transfected with the marked retroelements, with or without expression plasmids for human APOBEC3G (hA3G) and for the murine homologue APOBEC3 (mA3), respectively¹². G418^R cells were then cloned (Fig. 1c). In the absence of APOBEC3G, many clones were obtained with the

two retroelements (200–400 clones per 3×10^5 cells exposed to G418 selection). Both hA3G and mA3 markedly impaired retrotransposition of IAP and MusD, resulting in a fourfold and 50-fold decrease in the number of G418^R clones, respectively (Fig. 1d).

Non-LTR retrotransposons constitute the other broad family of autonomous retrotransposable elements in the mammalian genome^{1,4}. Non-LTR retrotransposons include L1 (LINE-1, or long interspersed nuclear elements-1), present at nearly 1 million copies in mouse or human genomes. About 80 human and 3,000 murine L1 elements are estimated to be full length and functional^{4,26,27} (Fig. 1a). The replicative cycle of non-LTR retrotransposons is different from that of endogenous retroviruses^{1,4}. Reverse transcription occurs within the nucleus, without the formation of identifiable VLPs. We examined whether APOBEC3G affects the L1 replicative cycle. In contrast with the results obtained with IAP and MusD, we found that two L1 elements (human L1Hs and murine L1Md) were insensitive to APOBEC3G activity in the retrotransposition assay (Fig. 1d).

The APOBEC group of proteins includes APOBEC1, the catalytic subunit of the apolipoprotein B mRNA editing complex, and APOBEC2, whose function is unknown^{12,17}. Murine APOBEC1 (mA1) and APOBEC2 (mA2) did not significantly inhibit retrotransposition of IAP, MusD and L1 elements (Fig. 1e).

We further asked whether APOBEC3G inhibits the retrotransposition of IAP and MusD elements by editing retroviral DNA. We sequenced 700-base-pair fragments of proviral DNA present in individual G418^R clones. Twenty-five clones were analysed for each retroelement (Fig. 2a). A comparison of the sequences revealed a large number of mutations only in the presence of APOBEC3G. The overall mutation rate was rather high, reaching a mean of seven mutations per virus-derived 700-base-pair fragment. This editing rate is most probably underestimated, because clones harbouring heavily mutated genomes (in which *neo* is hence rendered non-

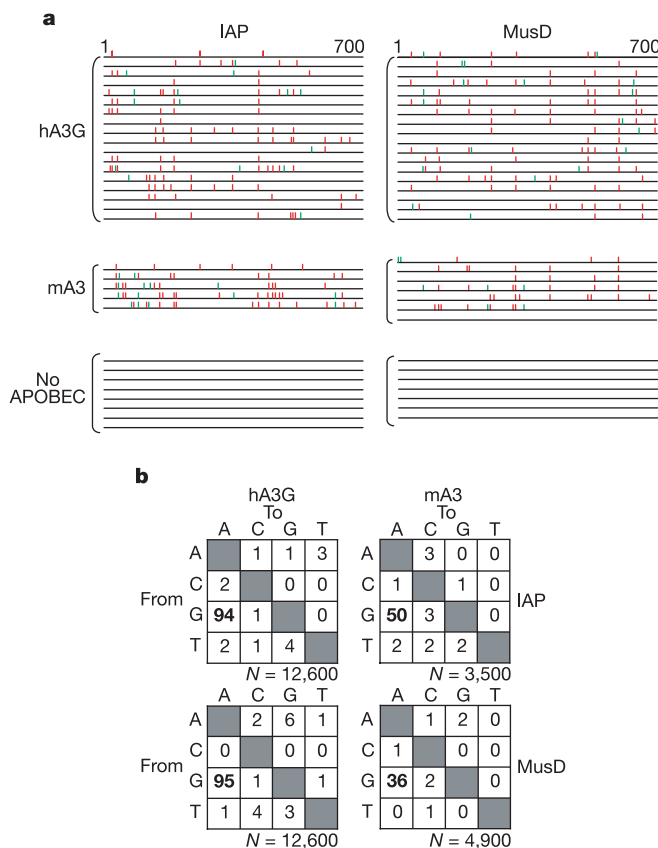


Figure 2 APOBEC3G promotes G-to-A hypermutation of endogenous retroviruses. **a**, Profiles of the position of the G-to-A mutations in *pol* sequences (700-base-pair fragment) of IAP and MusD retroelements, in the presence or absence of hA3G or mA3. Each line represents a sequence obtained from an independent G418^R clone. Mutations are colour coded: red, G-to-A; green, non-G-to-A. **b**, Nucleotide substitution preferences for the set of IAP and MusD *pol* mutations detected in the presence of hA3G or mA3. *N*, total number of bases sequenced.

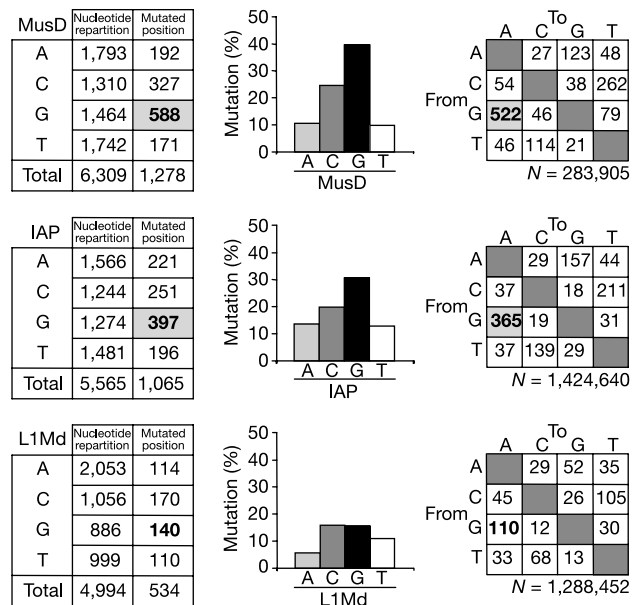
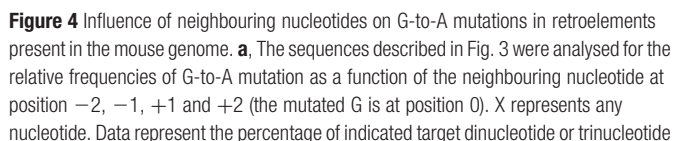


Figure 3 Distribution of nucleotide substitutions in MusD, IAP and L1Md retroelements present in the mouse genome. Endogenous sequences were extracted from the mouse genome database and aligned, and consensus sequences were generated. Left panels: nucleotide repartition in the consensus sequence and number of substitutions for each nucleotide in the analysed sequences. Middle panels: percentage of substitutions for each nucleotide. Right panels: nucleotide substitution preferences. *N*, total number of nucleotides analysed. Because a given residue could undergo different mutations, the sum of the mutations in the right panels is greater than the total number of mutated positions in the left panels.

We then undertook a characterization of the naturally occurring endogenous IAP, MusD and L1Md DNA sequences present in the mouse genome, and evaluated the levels of G-to-A substitutions in their multiple copies. Numerous mutations were present, reflecting the millions of years of genome evolution that have occurred since

Importantly, an overwhelming majority of MusD (44 of 45 analysed elements) and IAP (254 of 256 analysed elements) bear GXA-to-AXA substitutions. With MusD, the number of GXA-to-AXA mutations ranged from 1 to 72 per individual genome (with a mean of 15). Examples of MusD elements, in which sub-genomic regions were enriched in G-to-A substitutions, are shown in Fig. 4b. In these regions, G(G/A)A motifs were often mutated, similarly to the substitution profile observed in our cell-based experiments. It is noteworthy that these G-to-A mutations were distributed throughout the viral genome, with some local variations (Supplementary Fig. S2). Strikingly, in some MusD retroelements, up to 60% of all the observed single-nucleotide substitutions were GXA-to-AXA transitions (Supplementary Fig. S2). Similar overall results were observed with IAP: the number of GXA mutations ranged from 1 to



sequences bearing G-to-A mutations. **b**, Examples of G-to-A hypermutation in 12 MusD sequences present in the mouse genome. Two regions, corresponding to *gag* (upper panel) and *pol* (lower panel) regions are depicted. GXA trinucleotides are underlined in the consensus sequence. The localization of each retroelement in the mouse genome is reported in the legend to Supplementary Fig. S2.

33 per individual genome (with a mean of 4). Together, these observations are highly compatible with the existence of an APOBEC-like enzymatic activity *in vivo* that has affected the retrotransposition of MusD and, to a lesser extent, IAP retroelements.

Collectively, these studies reveal that APOBEC3G might protect the mammalian genome against the spread of some, but not all, retroelements by deaminating C residues during their retrotransposition. Consistent with a recent report²⁹ is our observation that L1 retroelements are not inhibited by APOBEC3G (a cytoplasmic enzyme), probably because reverse transcription of this class of retroelements occurs in the nucleus, in the absence of VLPs. It is also possible that APOBEC3G is unable to bind L1 components (proteins and/or nucleic acids) directly, precluding the expression of antiviral activity. In contrast, we show that APOBEC3G potentially inhibits the retrotransposition of IAP and MusD elements. The numerous mutations introduced by APOBEC are likely to inactivate these endogenous retroviruses, preventing further movement. MusD was more sensitive to APOBEC3G than IAP. It will be worth determining the origin of this phenomenon and identifying which viral components interact with APOBEC3G. Both murine and human enzymes were active against these two unrelated murine endogenous retroviruses. This strongly suggests that, in humans, the movement of some HERVs may be restricted through similar mechanisms. The anti-HIV effect of APOBEC3G is counteracted by Vif, which promotes degradation of the enzyme. It remains to be investigated whether endogenous retroviruses have developed methods of evading inactivation by APOBEC3G challenge.

The genome of endogenous retroviruses is generally methylated and not expressed in most somatic cells^{1,2,4}. These retroelements are hypomethylated and are hence active in both early embryos and germ cells^{8,14}, where APOBEC3G is also abundantly expressed^{12,16}. The germ line might therefore represent a natural site where APOBEC3G targets these endogenous retroviruses. Uncontrolled transposition or expression of mobile elements is obviously deleterious for the host, and this has been associated with pathological situations^{1,2,4,5}. In contrast, low levels of retrotransposition in germline cells, leading to the integration of inactive retrotransposons edited by APOBEC enzymes, might be beneficial and might contribute to host genome evolution. □

Methods

Plasmids

Plasmids coding for human and murine L1 (p220.CMV-L1.2Bneo^{TNF} and pCMV L1Md-Gf21neo^{TET}, respectively), IAP (pGL3-IAP92L23neo^{TNF}) and MusD (pCMVMus-6Dneo^{TNF}) retrotransposons have been described^{6,7,27,30}. The various APOBEC genes used in this study were cloned in pCDNA6 expression plasmid (Invitrogen) and sequenced. hA3G was obtained from the cDNA of human PBMCs. mA1, mA2 and mA3 were a gift from N. Landau¹². A plasmid expressing a defective hA3G gene (with a premature stop codon) was used as a negative control (no APOBEC points). All APOBEC proteins (which are epitope-tagged) were correctly expressed, as judged by western blot analysis (not shown).

Transfections and retrotransposition assays

Transfection and retrotransposition assays were described previously^{6,30}. In brief, cells seeded in 60-mm-diameter plates were transfected with 1.5 µg of neo-marked retrotransposons and 1.5 µg of APOBEC expression vectors by using the Lipofectamine Plus kit (Gibco BRL). Transfected cells were allowed to grow for 1 week and seeded at 3 × 10⁵ cells per 100-mm dish for G418 selection. At day 14 after selection, G418^R foci were either removed individually and plated, or fixed, stained and quantified. We ensured that the G418^R clones resulted from retrotransposition by verifying that the intron present in the initial neo gene was spliced out.

DNA analysis

Cellular DNA from individual G418^R clones was used as a template for PCR amplifications (Expand Long Template System; Roche). Thirty cycles of amplification were performed with sets of appropriated primers. PCR products were sequenced (Applied Biosystem sequencing kit). The DNA fragments sequenced corresponded to nucleotides 4224–4823 (IAP 92L23 clone), 4743–5342 and 5903–6502 (MusD AC124426 clone).

Mouse genome analysis

Endogenous retroelements were extracted from the mouse genome sequence database (Mouse GoldenPath mm5, May 2004) by using as a probe the three functional

retroelements studied in the cell-based retrotransposition assay. Sequences presenting the highest homology to the probes were aligned with ClustalW and Editsequence software. Forty-five MusD, 256 IAP and 258 L1Md elements were analysed, and consensus sequences were generated. The quantitative analysis of nucleotide substitutions within MusD, IAP and L1Md elements (results presented in Figs 3 and 4 and in the text) was performed using Excel software, on the coding domains of these retroelements (6,309 of 7,492 nucleotides for MusD, 5,565 of 7,084 nucleotides for IAP, 4,994 out of 7,521 nucleotides for L1Md). The locations of the analysed sequences within the mouse genome, and also the sequence alignments, are available from the authors on request.

Received 8 October; accepted 1 December 2004; doi:10.1038/nature03238.

- Boeke, J. D. & Stoye, J. in *Retroviruses* (eds Coffin, J. M., Hughes, S. H. & Varmus, H. E.) 343–436 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997).
- Bannert, N. & Kurth, R. Retroelements and the human genome: New perspectives on an old relation. *Proc. Natl Acad. Sci. USA* **101**(suppl. 2), 14572–14579 (2004).
- Johnson, W. E. & Coffin, J. M. Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl Acad. Sci. USA* **96**, 10254–10260 (1999).
- Kazanian, H. H. Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
- Lower, R. The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends Microbiol.* **7**, 350–356 (1999).
- Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G. & Heidmann, T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nature Genet.* **36**, 534–539 (2004).
- Ribet, D., Dewannieux, M. & Heidmann, T. An active murine transposon family pair: retrotransposition of 'master' MusD copies and ETn trans-mobilization. *Genome Res.* **14**, 2261–2267 (2004).
- Bourchis, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96–99 (2004).
- Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646–650 (2002).
- Mangeat, B. *et al.* Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424**, 99–103 (2003).
- Harris, R. S. *et al.* DNA deamination mediates innate immunity to retroviral infection. *Cell* **113**, 803–809 (2003).
- Mariani, R. *et al.* Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* **114**, 21–31 (2003).
- Heidmann, O. & Heidmann, T. Retrotransposition of a mouse IAP sequence tagged with an indicator gene. *Cell* **64**, 159–170 (1991).
- Dupressoir, A. & Heidmann, T. Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. *Mol. Cell. Biol.* **16**, 4495–4503 (1996).
- Baust, C. *et al.* Structure and expression of mobile ETnI retroelements and their coding-competent MusD relatives in the mouse. *J. Virol.* **77**, 11448–11458 (2003).
- Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79**, 285–296 (2002).
- Wedekind, J. E., Dance, G. S., Sowden, M. P. & Smith, H. C. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet.* **19**, 207–216 (2003).
- Lecossier, D., Bouchonnet, F., Clavel, F. & Hance, A. J. Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* **300**, 1112 (2003).
- Suspene, R. *et al.* APOBEC3G is a single-stranded DNA cytidine deaminase and functions independently of HIV reverse transcriptase. *Nucleic Acids Res.* **32**, 2421–2429 (2004).
- Yu, Q. *et al.* Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nature Struct. Mol. Biol.* **11**, 435–442 (2004).
- Wiegand, H. L., Doeble, B. P., Bogerd, H. P. & Cullen, B. R. A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. *EMBO J.* **23**, 2451–2458 (2004).
- Bishop, K. N., Holmes, R. K., Sheehy, A. M. & Malim, M. H. APOBEC-mediated editing of viral RNA. *Science* **305**, 645 (2004).
- Marin, M., Rose, K. M., Kozak, S. L. & Kabat, D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nature Med.* **9**, 1398–1403 (2003).
- Stopak, K., de Noronha, C., Yonemoto, W. & Greene, W. C. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* **12**, 591–601 (2003).
- Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* **2**, E275 (2004).
- Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
- Goodier, J. L., Ostertag, E. M., Du, K. & Kazanian, H. H. Jr. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.* **11**, 1677–1685 (2001).
- Liddament, M. T., Brown, W. L., Schumacher, A. J. & Harris, R. S. APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 *in vivo*. *Curr. Biol.* **14**, 1385–1391 (2004).
- Turelli, P., Vianin, S. & Trono, D. The innate antiretroviral factor APOBEC3G does not affect human LINE-1 retrotransposition in a cell culture assay. *J. Biol. Chem.* **279**, 43371–43373 (2004).
- Esnault, C., Casella, J. F. & Heidmann, T. A *Tetrahymena thermophila* ribozyme-based indicator gene to detect transposition of marked retroelements in mammalian cells. *Nucleic Acids Res.* **30**, e49 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Wain-Hobson for useful comments, N. Landau and the NIH AIDS reagents program for the gift of reagents. This work was supported by grants from the Institut Pasteur, the Institut Gustave Roussy, the ANRS, Sidaction, the Ligue Nationale contre le Cancer, the CNRS, INSERM and the European Community. F.D. is a fellow of Sidaction. Some relevant original papers have not been cited owing to space constraints.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.H. (heidmann@igr.fr) or O.S. (schwartz@pasteur.fr).

Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation

Marilyn G. Pray-Grant¹, Jeremy A. Daniel¹, David Schieltz², John R. Yates III³ & Patrick A. Grant¹

¹Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA

²Diversa Corporation, 4955 Directors Place, San Diego, California 92121, USA

³Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, USA

The specific post-translational modifications to histones influence many nuclear processes including gene regulation, DNA repair and replication¹. Recent studies have identified effector proteins that recognize patterns of histone modification and transduce their function in downstream processes². For example, histone acetyltransferases (HATs) have been shown to participate in many essential cellular processes, particularly those associated with activation of transcription³. Yeast SAGA (Spt-Ada-Gcn5 acetyltransferase) and SLIK (SAGA-like) are two highly homologous and conserved multi-subunit HAT complexes, which preferentially acetylate histones H3 and H2B and deubiquitinate histone H2B. Here we identify the chromatin remodelling protein Chd1 (chromo-ATPase/helicase-DNA binding domain 1) as a component of SAGA and SLIK. Our findings indicate that one of

the two chromodomains of Chd1 specifically interacts with the methylated lysine 4 mark on histone H3 that is associated with transcriptional activity. Furthermore, the SLIK complex shows enhanced acetylation of a methylated substrate and this activity is dependent upon a functional methyl-binding chromodomain, both *in vitro* and *in vivo*. Our study identifies the first chromodomain that recognizes methylated histone H3 (Lys 4) and possibly identifies a larger subfamily of chromodomain proteins with similar recognition properties.

Chromatin-modifying complexes regulate chromatin function by transferring different moieties to histones; specific biological processes are thought to be controlled through a histone 'code' that reflects distinct sequential and/or combinatorial patterns of modifications⁴. Patterns of histone acetylation and methylation are associated with organizing the transcriptional capacity of chromatin^{3,5}. SAGA and SLIK (also known as SALSA) are two Gcn5-dependent HAT complexes that contain many transcription-related proteins and the deubiquitinating enzyme Ubp8 and are involved in the regulation of RNA polymerase II-transcribed genes in yeast^{3,6-12}. The SAGA family of complexes has been remarkably conserved through evolution, suggesting that they play a pivotal role in transcription in all eukaryotes³.

Here we report the identification of Chd1 as a component of the *Saccharomyces cerevisiae* SLIK and SAGA complexes. Chd1 is the sole yeast member of the CHD family of proteins¹³, which contain two N-terminal chromodomains, a central helicase/ATPase domain and a DNA binding domain at the C terminus. Yeast Chd1 has been demonstrated to function in chromatin remodelling, gene expression and transcriptional elongation¹⁴⁻¹⁶. In our study, Chd1 was initially identified by mass spectrometry as a component of

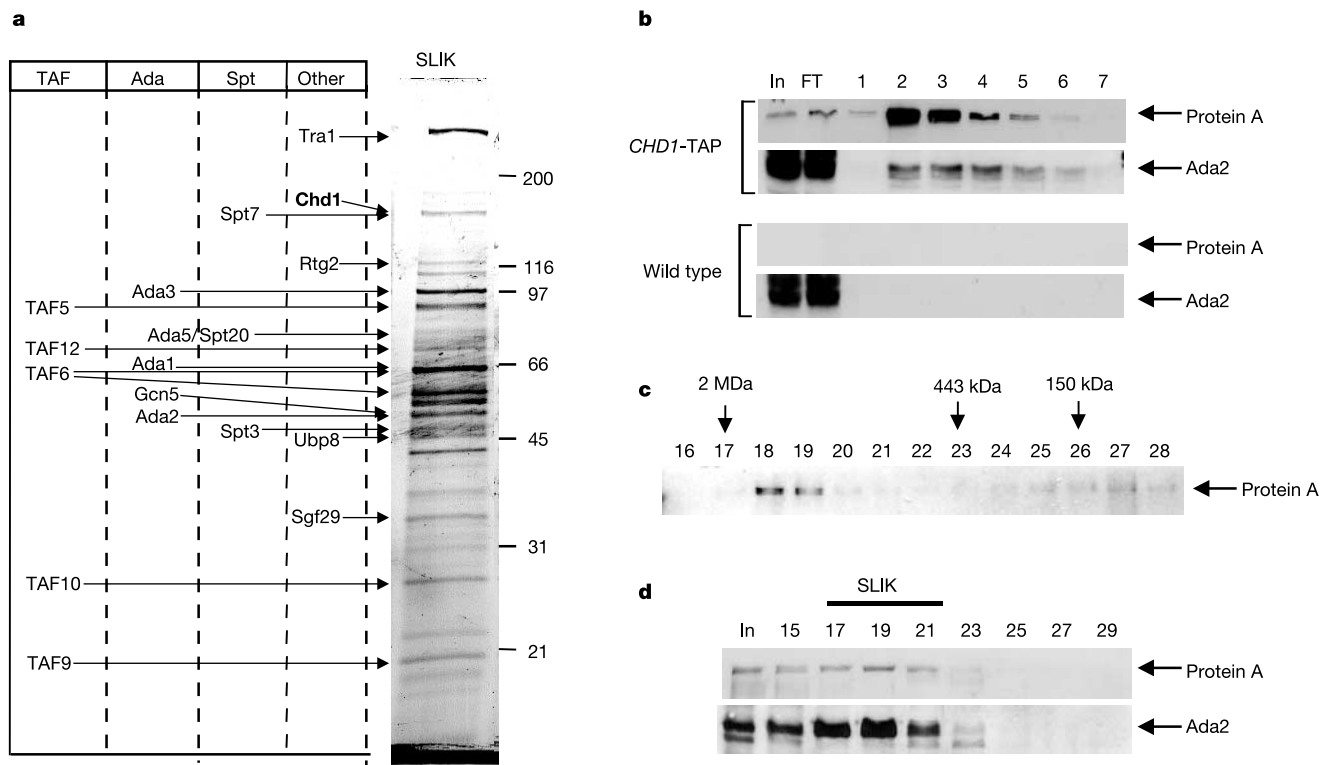


Figure 1 Identification of Chd1 as a component of SLIK. **a**, Silver-stained gels of highly purified yeast SLIK complex showing protein components that were identified by mass spectrometry. Proteins are grouped according to different classes of transcriptional regulators (TAF, Ada, Spt or others). Molecular masses in kDa are indicated. **b**, Whole cell extracts from *CHD1-TAP* and isogenic untagged strains were bound to and eluted from calmodulin beads, and the eluted fractions were analysed by western blotting. Shown are

western blots of input (In), flow-through (FT) and eluted fractions using Protein A (detecting Chd1-TAP) and Ada2 antibodies. **c**, Western blots of Superose 6 column-purified Chd1-TAP protein. The migrations of molecular weight standards are indicated with arrows. **d**, Western blots of purified SLIK complex with Ada2 and Protein A antibodies showing co-fractionation of Chd1 with SLIK complex.

purified native SLIK (Fig. 1a). To confirm association of Chd1 with other components of SLIK, whole cell extracts were prepared from a genomically integrated, C-terminal TAP-tagged *CHD1* strain and an untagged wild-type strain. Whole cell extracts were bound to calmodulin and then eluted with EGTA. Chd1 protein is detected by western blotting as a doublet that co-precipitates with Ada2 from the *CHD1*-TAP strain, but not from the untagged strain⁶ (Fig. 1b). Direct size-exclusion chromatography of the eluates reveals that Chd1 exists predominantly in a high molecular weight complex, consistent with its potential association with the SLIK complex (Fig. 1c). A portion of Chd1 also fractionates as a lower molecular weight complex and may represent the previously described monomer/dimer of 150–340 kDa¹⁴ or a complex with casein kinase II, which might function in phosphorylation of elongation factors¹⁵. To further characterize the association of Chd1 with subunits of SLIK, the SLIK complex was purified from the calmodulin eluates by sequential anion exchange and size-exclusion chromatography⁶. The fractions were analysed by western blotting and show that the Chd1-TAP protein co-fractionates with Ada2 as part of the 1.7 MDa SLIK complex (Fig. 1d), which may account for at least some of the implied functions of Chd1 in gene transcription¹⁶.

To determine the function of Chd1 in the context of SLIK, extracts were prepared from wild-type and *chd1Δ* yeast strains and fractionated over a Mono Q anion exchange column to isolate native SLIK and the related SAGA complex⁶. We found that Chd1 co-fractionated not only with SLIK, but also with SAGA prepared from wild-type extracts (Fig. 2a). Although their structural integrity appears to be intact, the HAT activity of both complexes on native histones is compromised in the *chd1Δ* strain (Fig. 2a). The activity of an unrelated histone H4–HAT complex⁶ eluting in the same fractions remains relatively unaltered. Introduction of a plasmid expressing an HA-tagged Chd1 protein into *chd1Δ* yeast largely restored the HAT activity of the SAGA and SLIK complexes (Fig. 2a). This indicates that Chd1 is a component of both SAGA and SLIK and is important for their functional HAT activity.

Chd1 contains two chromodomains, the functions of which are poorly understood. Chromodomains can play important roles in localizing proteins to chromatin¹⁷. The chromodomains of heterochromatin-associated protein 1 (HP1) and Polycomb (Pc), which function in gene repression and heterochromatin formation, have been demonstrated to recognize methylated lysines 9 and 27 respectively on histone H3; this determines their nuclear localization patterns^{18,19}. Thus far, the conserved chromodomains of HP1 and Pc are the only protein modules that have been shown to directly interact with any methylated histone. In contrast to HP1 and Pc, the Chd1 protein is associated (in a chromodomain-dependent fashion) with decompacted interphase chromatin in mammalian cells and transcriptionally active puffs and interbands in *Drosophila* polytene chromosomes^{20,21}. Furthermore, the chromodomains of yeast Chd1 are required for its role in transcription elongation¹⁶. Interestingly, di- and trimethylation of histone H3 Lys 4 are restricted to the transcribed regions in *Drosophila* and yeast^{22–24}, suggesting that Chd1 chromodomain function may be intimately linked to this chromatin modification.

Our finding that SAGA and SLIK complexes contain Chd1 prompted us to investigate whether SLIK has any acetyltransferase preference for a Lys 4-methylated substrate. Purified recombinant Gcn5 and native NuA3 and SLIK complexes were incubated with unmodified or Lys 4-dimethylated histone H3 peptides in *in vitro* HAT assays. The SLIK complex shows preferential acetylation of the methylated peptide compared to the unmodified peptide (Fig. 2b). This activity is not intrinsic to the catalytic subunit alone or universal to all HAT complexes, as neither Gcn5 nor the unrelated NuA3 complex distinguishes between the two substrates. Preferential acetylation by SLIK is not observed on a peptide dimethylated on H3 Lys 9 (data not shown), a modification that has only been reported in *S. pombe* and higher eukaryotes.

We wanted to determine whether the chromodomains of Chd1 are involved in SLIK recognition of a Lys 4-methylated histone H3 substrate. Chromodomains (CD) 1 and 2 of Chd1 were independently expressed and purified from bacteria as His-tagged proteins and incubated in binding assays with various biotinylated H3 amino-terminal peptides. CD1 was not found to co-precipitate with any peptide. However, CD2 interacted specifically with the Lys 4-dimethylated peptide and not the unmodified, acetylated Lys 9 or phosphorylated Ser 10 controls (Fig. 3b). The CD2 of Chd1 has not previously been predicted to interact with methylated histones, as it does not fall into the HP1 and Pc group of chromodomain-containing proteins²⁵. Hence, it is possible that CD2 of Chd1 represents a new subfamily of chromodomains (the CD2 subfamily) with the ability to recognize Lys 4-methylated histone H3 substrates. A PSI-BLAST sequence comparison of CD2 with the Swiss-Prot database identified numerous proteins with significant similarity to CD2 (Fig. 3a). All of these identified sequences are also members of the CHD superfamily, with the CD2 homology falling in the second of the two chromodomains.

The crystal structures of the chromodomains of both Pc and HP1 reveal that they contain three aromatic residues that form superimposable cages around the methyl-ammonium groups of the methylated lysine residues that they bind^{18,19,26,27}. Alignment of

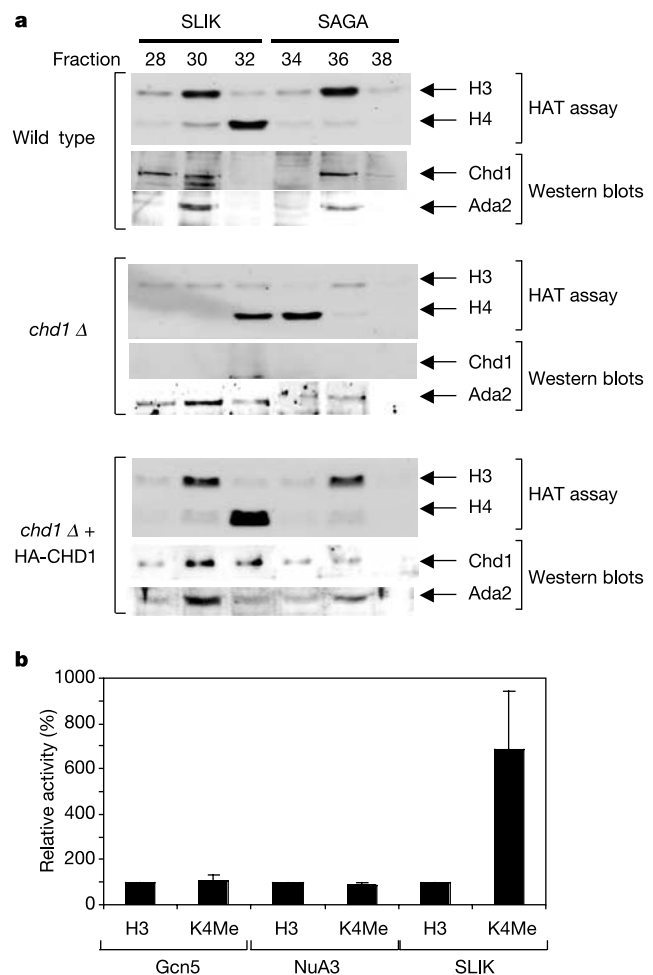


Figure 2 Defects in SAGA and SLIK HAT function in *chd1Δ* yeast. **a**, Whole cell extracts from wild-type and *chd1Δ* strains and *chd1Δ* yeast expressing HA-Chd1 were fractionated over a Mono Q column to separate SAGA and SLIK. Western blot analysis with Ada2 and Chd1 antibodies and fluorograms of HAT assays performed with the Mono Q fractions. **b**, HAT assays were performed using recombinant yeast Gcn5, purified NuA3 or SLIK complexes together with N-terminal tail peptides of unmodified histone H3 (H3) or Lys 4 dimethylated (K4Me) histone H3. Error bars show standard error.

the CD2 subfamily with the HP1 chromodomain shows that the first two aromatic amino acids are conserved (Fig. 3a). However, the third aromatic amino acid of HP1 aligns to an invariant leucine amino acid in the CD2 subfamily, which also has an invariant aromatic tyrosine residue two amino acids downstream (Fig. 3a). We postulated that if this tyrosine (Y316) of CD2 forms part of an aromatic cage recognizing methylated Lys 4, then mutation of this residue would perturb its interaction with methylated peptide. Targeted mutant forms of CD2 were generated to convert amino acids 314 and 316 to those found in HP1 (Fig. 3c). Mutation of Leu 314 (L314Y) did not affect binding of CD2 to the Lys 4 methyl peptide. However, mutation of the tyrosine (Y316E) disrupted stable CD2 binding to the peptide (Fig. 3d); the protein fractionated during the washing steps (data not shown). This suggests that Tyr 316 is required for stable association of Chd1 with histone H3 methylated at Lys 4 and that CD2 may form a similar aromatic cage structure to that of HP1 and Pc, critical for binding to methylated Lys 4.

Sequence alignment between CD1 and CD2 of Chd1 shows that CD1 contains the first two, but not the third aromatic residue within its chromodomain (Fig. 3c); this might explain why it does not bind to the methylated peptide (Fig. 3b). To test this hypothesis, mutations were generated in CD1 to give it features similar to CD2 (Fig. 3c). An E220L mutation of CD1 to insert the invariant leucine found in the CD2 group of chromodomains (Fig. 3a, c) did not significantly improve CD1 interaction with the methylated

peptide (Fig. 3e). However, introduction of an aromatic amino acid at position 222 (H222Y), analogous to position 316 of CD2, resulted in stable interaction between CD1 and the Lys 4-dimethylated peptide (Fig. 3e). This finding suggests that the KWxxLxY spacing of the second and third aromatic residues (which is conserved in the CD2 subfamily of CHD proteins) may form a potential methyl binding cage, important for methylated Lys 4 recognition. In contrast, the HP1 and Pc group of chromodomains, containing the sequence KWxxY/W (Fig. 3a) do not bind to H3 containing methylated Lys 4 (refs 18, 19, 25, 27).

To determine whether the chromodomains of Chd1 function in substrate recognition in the context of the SLIK complex, we used yeast strains expressing normal and mutant isoforms of HA-tagged Chd1 (ref. 16). Native SLIK complex was purified from wild-type yeast and yeast expressing Chd1 lacking both chromodomains (CDΔ) or bearing the point mutation Y316E in CD2. Western blotting showed that Chd1-CDΔ no longer interacts with the SLIK complex (Fig. 4a), but no gross changes in size (data not shown) or composition of the complex were observed (Fig. 4a). However, the Chd1-CDΔ protein is stably expressed in its truncated form in yeast cells (Fig. 4c). In contrast to Chd1-CDΔ, Chd1-Y316E remains associated with SLIK (Fig. 4a). Each of the three complexes was incubated with unmodified or Lys 4-methylated histone H3 peptides. The preferential acetylation of the methylated peptide observed with the wild-type complex is lost with SLIK complexes lacking Chd1 or bearing the CD2 mutation that abrogates methyl-

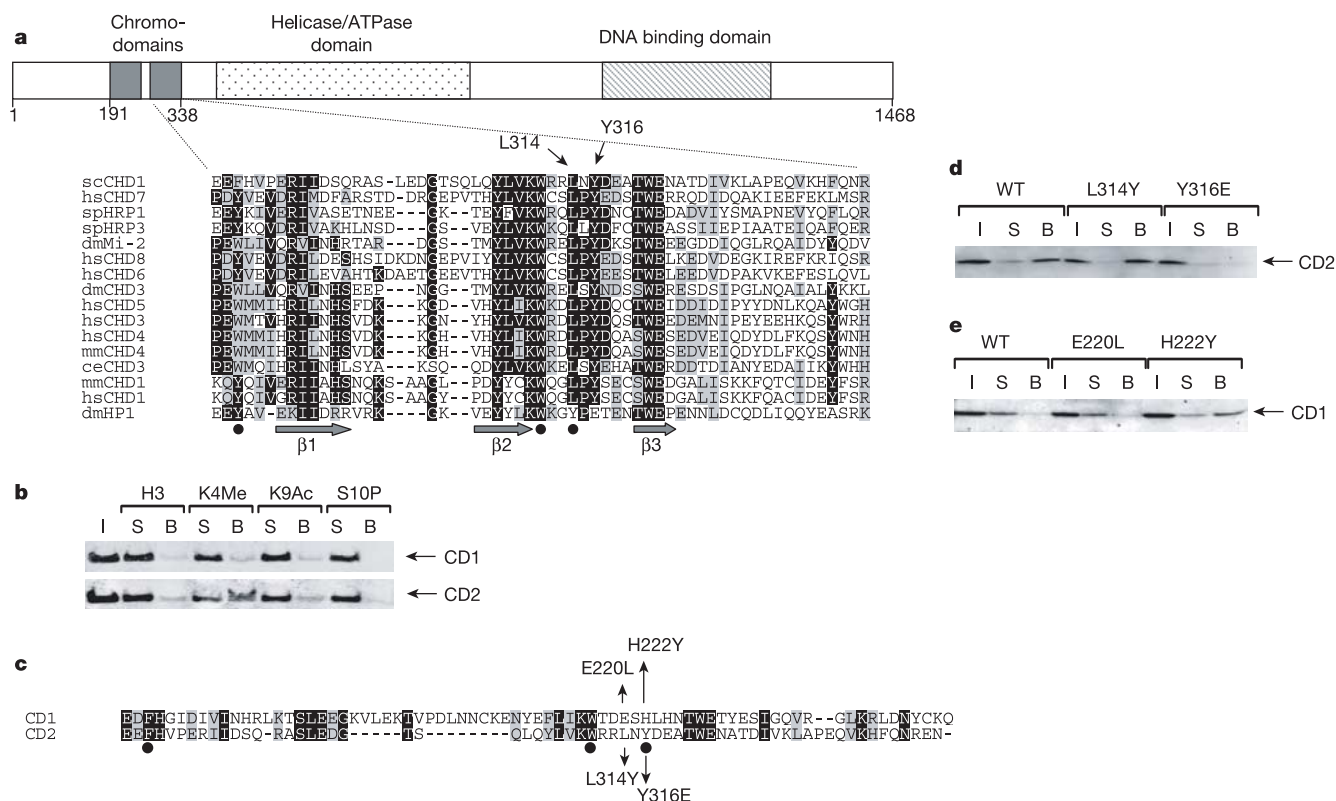


Figure 3 Chromodomain 2 of Chd1 mediates recognition of Lys 4-methylated histone H3. **a**, Illustration of the domain structure of yeast Chd1 and protein sequence alignment showing the CD2 of Chd1 and the most highly conserved homologues identified by PSI-BLAST multiple alignment of hits from the Swiss-Prot database. Black boxes highlight conserved residues and grey boxes highlight similarity between amino acids at given positions; proteins are ranked in order of their expectation value. *Drosophila* (dm) HP1 chromodomain is aligned below as a reference for the aromatic residues (filled circles) that form the methylated lysine binding cage and its β -strands (grey arrows). Positions of Chd1 residues L314 and Y316 are indicated. **b**, Purified His-tagged CD1 and CD2 of Chd1 were incubated with histone H3 N-terminal peptides, which were either unmodified (H3),

Lys 4-dimethylated (K4Me), Lys 9-acetylated (K9Ac) or Ser 10-phosphorylated (S10P). Inputs (I), supernatants (S) and bound fractions (B) from the reactions were analysed by western blotting using an anti-His antibody. **c**, ClustalW protein sequence alignment of CD1 and CD2 of Chd1. The aromatic residues in CD2 that are suggested to interact with methylated Lys 4 are shown (filled circles). Positions of point mutations made in CD1 (E220L, H222Y) and CD2 (L314Y, Y316E) are indicated with arrows. **d**, Wild-type (WT) and L314Y or Y316E mutant CD2 proteins were incubated with H3 Lys 4-dimethylated peptide and visualized as in **b**. **e**, Wild-type (WT) and E220L or H222Y mutant CD1 proteins were incubated with H3 Lys 4-dimethylated peptide and visualized as in **b**.

ated Lys 4 binding *in vitro* (Fig. 4b). The HAT activities of the complexes are not affected *per se*, because both the acetylation of unmodified peptides and the association of the catalytic subunit, Gcn5, are not significantly affected in either mutant complex (Fig. 4a, b).

To further validate our *in vitro* experiments, we used chromatin immunoprecipitation experiments to investigate whether Chd1 regulates the HAT activity of SLIK and SAGA *in vivo*. We and others have previously noted that when yeast are shifted into media containing galactose as a carbon source, induction of the *GAL* genes is accompanied by promoter-proximal histone H3 acetylation mediated by SAGA and SLIK^{6,11,28}. Yeast expressing wild-type Chd1, Chd1-CDΔ or Chd1-Y316E were grown in media containing raffinose or shifted for 30 min to media containing galactose. Chd1 expression was confirmed for each strain under the different conditions (Fig. 4c). Consistent with previous chromatin immunoprecipitation assays, acetylation of histone H3 accompanies *GAL* gene activation in yeast expressing wild-type Chd1 when they are induced in galactose media (Fig. 4d, e). However, yeast expressing Chd1-CDΔ or Chd1-Y316E did not display any significant increase in acetylation at the *GAL1-10* locus after galactose induction. At the *ACT1* locus, acetylation levels remain relatively normal in all strains under the conditions used. This indicates that a functional CD2 of Chd1 is required for full acetylation of a SAGA and SLIK-dependent promoter in native chromatin *in vivo*.

Transcriptionally competent euchromatin is methylated on

histone H3 at Lys 4, Lys 36 and Lys 79 (ref. 5). A recent study in *Drosophila* found a binary pattern of histone modifications among euchromatic genes, with active genes being hyperacetylated for histones H3 and H4 and hypermethylated at Lys 4 and Lys 79 of H3 (ref. 24). These results suggest a mechanism whereby enzymes responsible for histone methylation and acetylation can coordinate their enzymatic activity. Our results indicate that the SAGA and SLIK family of complexes may facilitate this set of events. Lys 4 methylation is regulated by the Set1 enzyme as part of the COMPASS complex and, like Chd1, is associated with the early stages of gene activation and with elongating RNA polymerase II⁵. Set1-mediated histone modification is also known to recruit the chromatin remodelling factor Isw1 (ref. 29) and to exclude the binding of repressor proteins from euchromatin⁵.

Methylation of Lys 4 at the *GAL1-10* locus is intimately tied to the ubiquitination status of Lys 123 on histone H2B, a modification which itself is regulated by SAGA and SLIK^{11,12}. Upon transcriptional activation at the *GAL1-10* upstream activating sequence (UAS), the transition to hypermethylated histone H3 is dependent upon Ubp8 function¹¹ and this methyl mark may in fact further stabilize SLIK and SAGA recruitment through Chd1 interaction. Along with Chd1 recognition of methylated H3, the bromodomain of Gcn5 binds acetylated histone peptides and is required for stable promoter occupancy by SAGA in the absence of transcriptional activators^{2,30}. Thus, an elaborate cascade of histone modifications is regulated by the SAGA and SLIK complexes, which in turn stabilize

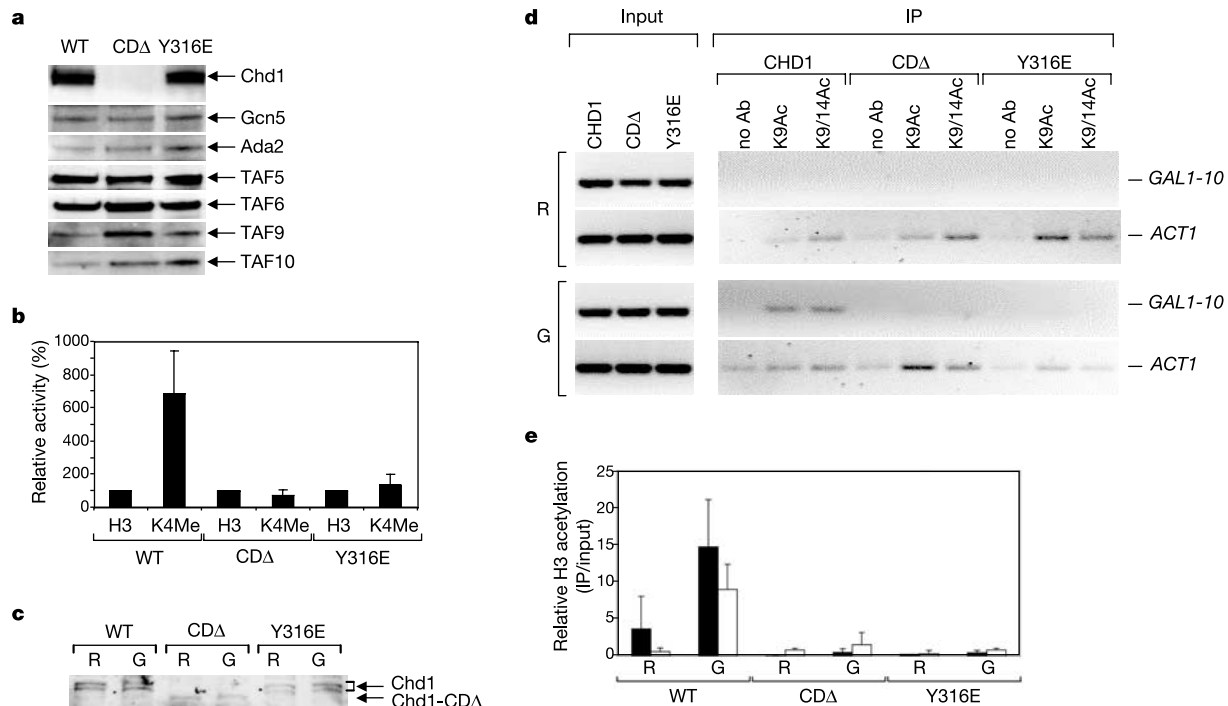


Figure 4 Chromodomain 2 of Chd1 regulates substrate recognition by SAGA and SLIK. **a**, SLIK complexes were purified from strains expressing wild-type HA-Chd1 (WT), HA-Chd1 lacking both chromodomains (CDΔ) and HA-Chd1 bearing the mutation Y316E. Peak Superose 6 fractions were analysed by western blotting with an anti-HA antibody (detecting Chd1) and various SLIK component antisera. **b**, HAT assays were performed using WT, CDΔ and Y316E SLIK complexes together with N-terminal tail peptides from unmodified histone H3 (H3) or Lys 4-dimethylated (K4Me) histone H3. Error bars represent standard error. **c**, Western blot analysis of yeast whole cell extracts prepared from yeast strains grown in raffinose (R) or galactose (G), using an anti-HA antibody. The positions of full-length Chd1 doublet in the WT and Y316E extracts and truncated Chd1 in the CDΔ are shown. **d**, Chromatin immunoprecipitation assays using yeast strains grown in

raffinose or galactose, using Lys 9-acetylated (K9Ac) or Lys 9/14-acetylated (K9/14) histone H3 antisera. A control immunoprecipitation was performed without antibody (no Ab). Shown are polymerase chain reaction (PCR) products using *GAL1-10* and *ACT1* UAS primers with input and immunoprecipitated (IP) DNA templates. **e**, Quantification of products from two independent sets of PCR reactions performed from chromatin immunoprecipitation assays as described above. The ratios of immunoprecipitated DNA to input DNA for the *GAL1-10* UAS were calculated and normalized to those for the *ACT1* UAS and are shown as relative amounts of immunoprecipitate per input signal (IP/input). Filled bars represent K9Ac and open bars represent K9/14Ac chromatin immunoprecipitations. Error bars show standard deviation.

and potentiate histone activity, contributing to the establishment of a transcriptionally competent epigenetic environment. □

Methods

Yeast strains and plasmids

SLIK used for mass spectrometric analysis was isolated from yeast strain CY396 as described⁶. Genomic TAP-tagging of *CHD1* was performed in the yeast strain BY4742 as described previously⁶. The *chd1Δ*, *chd1-CDΔ* and *chd1-Y316E* strains are derivatives of GHY279 (*his3Δ200 lys2-1288 ura3-52 leu2Δ1 chd1Δ :: HIS3 trp1Δ63*)¹⁶. GHY279 was transformed with plasmids pGH269 (encoding triple HA-tagged wild-type Chd1)¹⁶, pGH271 (encoding HA-Chd1 lacking both chromodomains (CDΔ))¹⁶ and pMG316 (encoding wild-type HA-Chd1 bearing the point mutation Y316E).

Purification of SAGA and SLIK complexes and mass spectrometry

The highly purified SLIK complex used for silver staining and mass spectrometric procedures has been described previously⁶. SAGA and SLIK were also partially purified from 4 litres of wild-type, *chd1Δ* and *chd1Δ* + HA-*CHD1* strains grown to mid-log phase in YPD (yeast extract-peptone-dextrose). Whole cell extracts were bound to a Mono Q HR 16/10 column (Amersham) and fractions were eluted with a 500 ml linear gradient from 100 to 500 mM NaCl. Peak fractions were determined by HAT assays and by western blotting. Whole cell extracts from 2 litres each of wild-type and *CHD1*-TAP strains were prepared in the same manner and bound to 200 μl calmodulin resin (Stratagene) and eluted as described previously⁶. Peak fractions were pooled and loaded onto sequential Mono Q HR 5/5 and Superose 6 PC 3.2/30 columns (Amersham) to purify SLIK, as described⁶. SLIK was also partially purified from 4 litres of cultured wild-type, *chd1Δ* + HA-*chd1-CDΔ* and *chd1Δ* + *chd1-Y316E* strains, using Mono Q HR 16/10, Mono Q HR 5/5 and Superose 6 HR 10/30 columns (Amersham). Nickel-agarose was avoided in purification steps, as this disrupts stable Chd1 interaction with SAGA and SLIK (data not shown).

Western blot analysis and HAT assays

Protein A antibody (Sigma) was used to detect Chd1-TAP (Fig. 1). Chd1 and Chd1 derivatives were detected with an anti-Chd1 antibody (Fig. 2); an anti-HA antibody (Roche) was used to detect HA-Chd1, HA-Chd1-CDΔ and HA-Chd1-Y316E in Fig. 4. Anti-His antibody (Santa Cruz) was used to detect the chromodomain polypeptides (Fig. 3). HAT assays were performed with HeLa cell core histones and H3 N-terminal Lys 4-dimethylated or unmodified peptides (residues 1–20) as described⁶.

Chromodomain expression and binding assays

Sequences encoding Chd1 CD1 (174–248) and CD2 (282–343) and derivatives (Fig. 3c) were cloned into a PET16b vector (Novagen) and expressed from BL21pLysS bacterial cells as 10-His N-terminally tagged polypeptides under native conditions, using standard procedures. CD proteins were purified using Nickel-NTA agarose (Qiagen), dialysed to remove imidazole then concentrated. For binding assays, approximately 500 ng CD polypeptide was incubated with 2 μg biotinylated dimethylated Lys 4, acetylated Lys 9, phosphorylated Ser 10 or unmodified histone H3 N-terminal peptides 1–21 (Upstate) and 30 μl streptavidin agarose (Upstate), in binding buffer (150 mM NaCl, 50 mM HEPES, 0.1% Tween, 10% glycerol, 2 μg ml⁻¹ pepstatin, 2 μg ml⁻¹ leupeptin, 1 mM phenylmethylsulphonyl fluoride (PMSF) and 0.5 mM dithiothreitol (DTT)), for 2 h at 4 °C with rotation. The beads were collected by centrifugation and washed twice with binding buffer. Inputs, supernatants and beads were run on 4–20% SDS-PAGE gels and immunoblotted with an anti-His antibody.

Chromatin immunoprecipitation assays

Chromatin preparations and immunoprecipitations were performed largely as described previously¹¹. Cultures (100 ml) of the *CHD1*, *chd1-CDΔ* and *chd1-Y316E* yeast strains were grown in SC medium containing 2% raffinose to an absorbance at 600 nm of ~1.0, split into two 50-ml cultures then reseeded in SC-*raffinose* medium or SC medium containing 2% galactose for an additional 30 min. Chromatin from whole cell extracts was immunoprecipitated with 4 μl αH3 K9Ac or 2 μl αH3 K9/14Ac antibodies (Upstate). DNA was amplified using oligonucleotide primers specific for the *GAL1* UAS and *ACT1* and was visualized by ethidium bromide staining of a 2% agarose gel.

Protein alignments

Homologues of the second chromodomain of yeast Chd1 were identified after two iterations of a position-specific iterated (PSI) BLAST alignment from the Swiss-Prot database, with an *E* value <0.001. The sequences were subsequently aligned using ClustalW together with the *Drosophila* HP1 chromodomain.

Received 24 September; accepted 29 November 2004; doi:10.1038/nature03242.

Published online 12 January 2005.

1. Iizuka, M. & Smith, M. M. Functional consequences of histone modifications. *Curr. Opin. Genet. Dev.* **13**, 154–160 (2003).

2. Bottomley, M. J. Structures of protein domains that create or recognize histone modifications. *EMBO Rep.* **5**, 464–469 (2004).
3. Torok, M. S. & Grant, P. A. Histone acetyltransferase proteins contribute to transcriptional processes at multiple levels. *Adv. Protein Chem.* **67**, 181–199 (2004).
4. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
5. Sims, R. J., Nishioka, K. & Reinberg, D. Histone lysine methylation: a signature for chromatin function. *Trends Genet.* **19**, 629–639 (2003).
6. Pray-Grant, M. G. *et al.* The novel SLIK histone acetyltransferase complex functions in the yeast retrograde response pathway. *Mol. Cell. Biol.* **22**, 8774–8786 (2002).
7. Grant, P. A. *et al.* Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. *Genes Dev.* **11**, 1640–1650 (1997).
8. Sterner, D. E., Belotserkovskaya, R. & Berger, S. L. SAGA, a variant of yeast SAGA, contains truncated Spt7, which correlates with activated transcription. *Proc. Natl Acad. Sci. USA* **99**, 11622–11627 (2002).
9. Lee, T. I. *et al.* Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature* **405**, 701–704 (2000).
10. Huisinga, K. L. & Pugh, B. F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in *Saccharomyces cerevisiae*. *Mol. Cell* **13**, 573–585 (2004).
11. Daniel, J. A. *et al.* Deubiquitination of histone H2B by a yeast acetyltransferase complex regulates transcription. *J. Biol. Chem.* **279**, 1867–1871 (2004).
12. Henry, K. W. *et al.* Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. *Genes Dev.* **17**, 2648–2663 (2003).
13. Woodage, T., Basrai, M. A., Baxevanis, A. D., Hieter, P. & Collins, F. S. Characterization of the CHD family of proteins. *Proc. Natl Acad. Sci. USA* **94**, 11472–11477 (1997).
14. Tran, H. G., Steger, D. J., Iyer, V. R. & Johnson, A. D. The chromo domain protein chd1p from budding yeast is an ATP-dependent chromatin-modifying factor. *EMBO J.* **19**, 2323–2331 (2000).
15. Krogan, N. J. *et al.* RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell. Biol.* **22**, 6979–6992 (2002).
16. Simic, R. *et al.* Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. *EMBO J.* **22**, 1846–1856 (2003).
17. Brehm, A., Tufteland, K. R., Aasland, R. & Becker, P. B. The many colours of chromodomains. *Bioessays* **26**, 133–140 (2004).
18. Fischle, W. *et al.* Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev.* **17**, 1870–1881 (2003).
19. Min, J., Zhang, Y. & Xu, R. M. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev.* **17**, 1823–1828 (2003).
20. Stokes, D. G., Tartof, K. D. & Perry, R. P. CHD1 is concentrated in interbands and puffed regions of *Drosophila* polytene chromosomes. *Proc. Natl Acad. Sci. USA* **93**, 7137–7142 (1996).
21. Kelley, D. E., Stokes, D. G. & Perry, R. P. CHD1 interacts with SSRP1 and depends on both its chromodomain and its ATPase/helicase-like domain for proper association with chromatin. *Chromosoma* **108**, 10–25 (1999).
22. Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl Acad. Sci. USA* **99**, 8695–8700 (2002).
23. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002).
24. Schubeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
25. Jacobs, S. A. *et al.* Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J.* **20**, 5232–5241 (2001).
26. Nielsen, P. R. *et al.* Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* **416**, 103–107 (2002).
27. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080–2083 (2002).
28. Krebs, J. E., Fry, C. J., Samuels, M. L. & Peterson, C. L. Global role for chromatin remodeling enzymes in mitotic gene expression. *Cell* **102**, 587–598 (2000).
29. Santos-Rosa, H. *et al.* Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol. Cell* **12**, 1325–1332 (2003).
30. Hassan, A. H. *et al.* Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* **111**, 369–379 (2002).

Acknowledgements We thank D. Allis and G. Hartzog for peptides, yeast strains and plasmids and for discussions regarding the design and interpretation of these experiments. We also thank J. Workman, J. Reese and S. Berger for antisera and S. Khorasanizadeh for advice on expression of the Chd1 chromodomains. P.A.G. was the recipient of a Burroughs Wellcome Fund Career Award in Biomedical Sciences. J.A.D. is supported by a predoctoral cancer training grant from the University of Virginia Cancer Center. This work is supported by the NCCR Yeast Center and an NIDDK grant to P.A.G.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.A.G. (email: pag9n@virginia.edu).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Andy Douglas (4975)
Nils Moeller (4953)
Irene Vigila-Atton (4944)

Scandinavia/ Spain/ Portugal:
Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequinot (4974)

Advertising Production

Manager: Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:
Tom Hancock

Naturejobs online production:
Niamh Shields

European Satellite Office

**Germany/ Austria/ Italy/
The Netherlands/ Belgium:**

Patrick Phelan
Tel +49 89 54 90 57 11
Fax +49 89 54 90 57 20
e-mail: p.phelan@nature.com
Sharon de Weert
Tel +44 (0) 20 7843 4970
e-mail: s.deveert@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatomachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
Asia-Pacific Sales Director:
Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Central stations

Are you a PhD who is tired of fighting for work in an over-crowded market? Well, you might consider moving to Estonia. Rein Vaikmäe, a scientific adviser to the Estonian government, estimates that his nation needs to produce about 200 PhDs a year to grow its science base. It currently averages about 80 — and many of them leave the country.

Heading for Estonia, or one of the other new members of the European Union (EU), for scientific employment might not be so far-fetched. Several schemes are pouring money into these countries to repatriate its best young scientists who left during tougher times (see page 440). Mart Ustav, who now heads the microbiology and virology department at the University of Tartu in Estonia, recalls the old days all too clearly. “You couldn’t travel. You couldn’t communicate. You were behind the barbed wire,” he says.

Times have changed drastically. Repatriation and funding from philanthropies have seeded a critical mass in select scientific disciplines throughout central Europe — often planted in classical scientific training the Soviet era left behind. What is missing, say Vaikmäe, Ustav and other scientists in the region, is infrastructure and national funding.

But the next generation of scientists is already finding ways to work with what they’ve got. Karolina Tkaczuk, a graduate student at the Technical University of Lodz in Poland, made her research choice based on her country’s strengths and limitations (see page 442). She uses the hour-long train commute between Lodz and Warsaw, where she is doing her research training, to study and think. If more investment continues to pour into central Europe — from individual nations, as well as from philanthropies and the EU — then Western scientists might be willing to come along for the ride.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

The rise of Central Europe p440

CAREER VIEW

Recruiters & Academia

Top 10 tips for success in graduate school

Graduate Journal

Seeking sources of information

Movers

Ingelise Saunders

p442

WWW.NATUREJOBS.COM

Career centre

Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Homeward bound

European Union

Fifteen years after the collapse of the Soviet Union, Andres Metspalu still needs, politely, to enlighten many callers that Estonia is no longer a Moscow-steered country. The small Baltic republic is one of eight central European countries that joined the European Union (EU) last year. "But many people in western Europe and the United States still think we speak Russian," says the geneticist at the University of Tartu.

Metspalu is the director of the Estonian Genome Project and one of the country's most prominent researchers. Estonia's small size — it has a population of just 1.4 million people — allowed it to transform its science system in the 1990s more rapidly and efficiently than most other countries in the region. And Metspalu's genome project has further established Estonia on the worldwide map of research. But despite these achievements, even Estonia is struggling to cope with the exodus of many of its best young scientists to the West.

EU membership means that scientists in poorly funded central European countries can easily relocate to countries offering more research money and modern infrastructure. In 2004, some 15% of the EU's Marie Curie fellowships, which facilitate mobility of young scientists in Europe, were awarded to scientists leaving central Europe, whereas only about 1% went to a university, research institute or industry lab in the new EU countries. Although some labs in central Europe have benefited from European Commission return grants, given to former Marie Curie fellows and postdocs returning home after five years or more abroad, they still haven't made up for the net loss.

This flow of scientific talent into western labs has caused European and US research agencies to set up special programmes aimed at central European scientists who would like to go home, but have been deterred by low salaries and lack of modern facilities. The UK Wellcome Trust, the US Howard Hughes Medical Institute (HHMI), the European Commission and many national grant-giving agencies all run schemes aimed at helping postdocs to go back to central Europe. The next few years should see growing opportunities for central Europeans to return home, or for young scientists in new EU countries to begin a scientific career. But without more funding from individual countries, the long-term picture remains unclear.

The HHMI began funding scientists in central Europe in 1995, when the dissipation of the Soviet Union redrew political boundaries and left many

talented scientists without infrastructure or support. "We felt that these countries had a strong scientific tradition but were seriously short of funding," says Jill Conley, HHMI international programme director. So the institution initiated the International Research Scholars programme, which now funds 132 scientists in 29 countries, including 48 in central European and Baltic countries.

INSTALLING HELP

The HHMI is now focusing specifically on investigators in the Baltic states, central and eastern Europe, Russia and Ukraine. This December the institution will award US\$17.5 million to 40–50 investigators over five years. It has also teamed up with the European Molecular Biology Organization (EMBO) in Heidelberg, Germany, to fund 'installation grants' that would help young investigators in central Europe set up their first lab (see *Nature* **432**, 262; 2004).

The impact of such programmes is probably felt most not by primary investigators — many of whom would have received funding in a western country if they chose to take advantage of the EU's mobility — but by the home-grown students and postdocs who make up their lab. Tamas Freund, director of the Institute of Experimental Medicine at the Hungarian Academy of Sciences, receives several offers a year to leave Hungary for a western university job. His HHMI grant allowed him to stay put — but also meant he could retain several students and staff scientists for a five-year stint.

Mart Ustav, head of the microbiology and virology department at the University of Tartu, had a good job at Cold Spring Harbor Laboratory before his native Estonia became independent

of the Soviet Union. When he returned, Ustav used money from the Soros Foundation to jump-start the department. Now he receives HHMI funds for his group and another departmental colleague has Wellcome Trust money, allowing the department to fund three groups — a total of about 60 young scientists.

In at least one case, dual awards to single investigators are helping to establish critical mass in individual labs. Priit Kogerman, a cancer researcher at the National Institute of Chemical Physics and Biophysics in Tallinn, Estonia, is one of 23 young researchers in the region who have been awarded a Wellcome Trust International Senior Research Fellowship (ISRF), for outstanding mid-career scientists



Andres Metspalu (top) and Tamas Freund hope to retain gifted scientists in eastern Europe.

M. LEEGO

K. KALLBERG/HHMI



wishing to return to or remain in their home countries. Kogerman is also one of 48 scientists from the Baltics, central and eastern Europe to receive funds through HHMI's International Research Scholars programme.

A former postdoc at the Karolinska Institute in Stockholm, Sweden, Kogerman says that he planned to return ever since he left Estonia in 1993. The ISRF and HHMI money helped him build a high-profile group at home. "Things have improved here quite a bit since the days of hardship in the early 1990s," he says. "When I first started here, we worked with all the 20-year-old Soviet centrifuges and apparatus."

FLEXIBLE APPROACH

The ISRF scheme was set up in 2000, following a review of the quality of science and research communities in the then EU accession countries. The individual, highly flexible grants, which include a competitive salary, are available in five fields but, within Europe, are limited to Estonia, Poland, the Czech Republic and Hungary: the four countries with the strongest science bases in the region. The programme now funds 23 fellows — including six in Estonia and ten in Hungary — who receive some €750,000 (US\$980,000) for five years each.

"This is not an aid programme," says Hans Hagen, the Wellcome Trust's programme officer for basic immunology and infectious diseases. "We wanted to create some 'honey pots' to show that it is possible to do top-class research in central Europe, and to have a bit of leverage with local funding mechanisms."

One goal of extending the programme, which has been available in Australia, New Zealand, South Africa and India, is to establish a culture of postdoc research in the new EU countries. Lack of fixed-term positions and low mobility among scientists are common

legacies of the academic systems in central Europe and are obstacles to high-quality science, says Hagen.

Those who have returned agree that it is crucial to gain experience in the West. "I rate my students and postdocs very highly, but I do tell them they should go abroad after a few years to propel their careers," says Gabor Tamas, a Wellcome Trust-funded neuroscientist at the University of Szeged in Hungary.

Like all Wellcome Trust fellows, Tamas has a strong research record. He used part of his grant money to equip his new lab with super-modern apparatus, such as a prototype piece of equipment for recording signals from multiple neurons at the same time.

Thus, he says, he can easily maintain the same quality of work he became used to at his former lab at the University of Oxford, UK. "Don't talk to me about 'critical mass'," he says. "We're in the age of the Internet, and you no longer need to be in Oxford or Cambridge to do good science."

Most of those who have returned say they want to stay permanently. "If I find a follow-up grant I will certainly stay. From a scientific point of view there is no reason to leave," says Zdena Palkova, a yeast-cell biologist at the Charles University of Prague, Czech Republic, who in 2001 won a grant from the EMBO Young Investigator Programme, and in 2003 an additional HHMI grant.

But with national research funding in the new EU countries still low, western grants will be crucial, particularly in highly competitive fields such as molecular biology or neuroscience. "I see no chance in the next few years that Hungary, or any other country in the region, can compete without western resources," says Tamas.

Quirin Schiermeier is *Nature's* German correspondent; Paul Smaglik is editor of *Naturejobs*.



External grants have helped Priit Kogerman build a high-profile group in his native Estonia.

Web links

Wellcome Trust
www.wellcome.ac.uk
 Howard Hughes Medical Institute
www.hhmi.org
 EMBO Young Investigator Programme
www.embo.org/projects/yip/index.html
 Marie Curie fellowships
www.cordis.lu/improving/fellowships/home.htm
 EU mobility portal
europa.eu.int/eracareers/index_en.cfm
 EU science indicators
europa.eu.int/comm/research/era/pdf/indicators/ind_kf0304.pdf

GRADUATE JOURNAL

Seeking sources of information

Many central European students underestimate their scientific value. We think we are not educated enough because our labs lack modern equipment or do not use the most recent techniques. That is why we search for additional sources of learning.

I found a way of measuring myself against international students when I applied for a traineeship at one of Poland's most renowned institutes: the Laboratory of Bioinformatics and Protein Engineering at the International Institute of Molecular and Cell Biology in Warsaw (IIMCB).

I chose bioinformatics because it is a relatively young science, not yet as popular in Poland as in other countries. It is a good fit for central European scientists because it does not limit you to experiments in the wet lab, where we lack equipment, and it lets you teach yourself, to some extent.

Studying at the IIMCB will help me improve skills I gained during my master's project, when I worked on protein homologue modelling. I had to learn how to do this on my own because nobody at my university worked on this kind of bioinformatics problem. I've always been ambitious, so I anticipate that learning from the best in Poland will help me be as good a scientist as those in cutting-edge Western labs. ■

Karolina Tkaczuk is a graduate student at the Technical University of Lodz, Poland.

RECRUITERS & ACADEMIA

Top 10 tips for success in graduate school

Ensure that the choices you make enhance both your work and your career prospects.

10 Starting out Don't take classes as an undergraduate that you will take as a post-grad. Go to college to get a broad education — go to graduate school to earn a union card (or PhD).

9 Choosing an adviser Your adviser is like your spouse. Choose one you respect as a scientist and as a person. The department is your family-in-law. Pick one that has high standards and will challenge you.

8 Know thyself Recognize your strengths, weaknesses, likes and dislikes. Successful people emphasize their strengths, minimize their weaknesses.

7 Write a thesis you can be proud of You will probably

only ever write one doctoral thesis — make it worth the effort.

6 Do good science One cannot over-emphasize the importance of hard-nosed, well-controlled science. When things go badly, it is the only thing that will get you out of trouble. When things go well, it will stop you doing something silly. Be as critical of your data as you are of other people's.

5 The best scientists are artists You will be judged by the quality of your questions as well as the technical excellence you use to address them. Crafting good hypotheses requires imagination, insight and the ability to see what others do not.

4 No guts, no glory You must be brave to be creative. Develop the self-confidence to risk failure, to ask big questions and to aim high. Also, understand when you have been lucky.

If you do not, there is no point in being lucky.

3 Data are royalty Only data are sacrosanct: not hypotheses or beliefs. Critically analyse your data. Don't just look for what you want. Some of the most interesting results are those we did not expect. You can't reject data because you don't like them or they confuse you.

2 Have fun Success takes a lot of work and dedication: you have to love and enjoy the process. Try to make your hobby your career.

1 Remember you are searching for truth Your objective is to join the community of scholars; the sole function of scholarship is to find the truth. Truth is elusive, but the scientific method gives us the most objective criteria available to judge 'truth'. ■

Primal de Lanerolle is a professor of physiology at the University of Illinois at Chicago.

MOVERS Ingelise Saunders, chief executive, ACE BioSciences, Odense, Denmark



Postgraduate benchwork was "boring", says Ingelise Saunders — so she made a change that set her on the path to success. Starting as a medical sales representative, she went on to work her way up the drug-company ranks to chief executive. "I didn't enjoy the lab work at all," Saunders says. "It was quite clear to me that I needed to get out and work with people."

She found that one of the best ways to work with people — and learn the pharmaceuticals business — was as a product representative, first for Wyeth Ayerst, then Schering Plough and finally Glaxo. She found that experience

invaluable. "What you learn from being a rep — even if you are only repping for one or two years — is how to work with customers and understand people's needs," Saunders says.

Repping is a maligned and often-overlooked career path for scientists, she adds. But it can lead to other opportunities. She has seen colleagues who started as sales reps go on to management or marketing positions — or even back into academia (see *Nature* **430**, 486–487; 2004).

The sales background helped prepare her to make the switch into marketing management. She benefited from a combination of good mentoring and formal management training. "I was fortunate enough to have some very good managers who coached me and helped me," Saunders says.

But after years of working her way to the upper ranks of the large drug companies, she decided in 2001 to

downsize. "I got tired of the big organizations with all the politics and all the bureaucracy," she says.

At Celltech, she "thoroughly enjoyed" being more involved with every aspect of a company. She left after the London-based biotech company was bought by UCB. Her newest job, transforming ACE BioSciences from a research-based organization focused on infectious diseases into a product-based company, will be a big challenge — especially as the company has only a couple of dozen employees. She is rolling up her sleeves and pitching in with the sort of admin tasks she used to delegate.

Saunders sees the move to ACE as a return to her roots in several ways. First, it has allowed her to move back to her native Denmark to head the Odense-based company. And second, she will once again be driving to make sales — but this time to investors. Just like a rep, but with higher stakes. ■

CV **2001–04:** Director, European operations rising to chief executive, Celltech, London.
1986–2001: Area manager rising to chief executive, NovoNordisk, Denmark.
1983–86: Sales and marketing director, Glaxo, London.
1980–82: Sales manager, Schering Plough, London.